Digitale Bibliothek : Konzeption und Implementierung mit der *Greenstone Digital Library Software*

Diplomarbeit
im Fach Digitale Bibliothek,
Studiengang Wissenschaftliche Bibliotheken
der
Fachhochschule Stuttgart –
Hochschule der Medien,
Fachbereich Information und Kommunikation

Florian Engster

Erstprüferin Prof. Margarete Payer, Stuttgart
Zweitprüfer Dipl.-Bibl. Stefan Wolf, Konstanz

Bearbeitungszeitraum: 15. Juni 2002 bis 15. Oktober 2002

Stuttgart, Oktober 2002

Kurzreferat

Die Idee der digitalen Bibliotheken blickt auf eine lange Tradition zurück, doch fehlt es bislang an konkreten konzeptionellen Überlegungen zur Einrichtung eines solchen Dienstes, der seinem Namen auch gerecht wird.

Es wird ein Katalog von Kriterien aufgestellt, die als Maßstab für eine digitale Bibliothek gelten sollen. Auf Basis der *Greenstone Digital Library Software* wird eine Modell implementiert, das diese Kriterien berücksichtigen soll.

Schlagwörter (SWD)

s.Elektronische Bibliothek ; s.Planungskonzept ; s.Errichtung

Schlagwörter (INFODATA)

Bibliothek ; Informationsversorgung ; Digital ; Planung

Notation (ACM CCS 98)

H.3.7 ; H.1.1

Abstract

The idea of digital libraries emerges from a long tradition but lacks of definitive planning for the creation of such a service, which satisfies it's name.

This writing establisches a catalogue of criteria to serve as a template for the implementation of a digital library. The *Greenstone Digital Library Software* will serve as the kernel for a model, which tries to match these criteria.

Subject Terms (INFODATA)

Library ; Information Supply ; Digital ; Planning

Classification (ACM CCS 98)

H.3.7 ; H.1.1

Vorwort

Digitale Bibliotheken haben mich durch die Möglichkeit fasziniert, Information effizient und einfach zu verwalten und einen schnellen Zugriff auf große Mengen von Daten zu ermöglichen. Zudem besteht in der gemeinsamen Nutzung bibliothekarischer Ansätze und Informationstechnik ein erhebliches Potential.

Die vorliegende Arbeit ist das Ergebnis einer eingehenden Beschäftigung mit dem Thema der informationstechnisch gestützten Informationsvermittlung. Vor allen anderen Dingen interessierte mich aber die konzeptionelle Seite digitaler Bibliotheken, die vor den schier grenzenlosen Möglichkeiten der Technik in meinen Augen immer wieder zu kurz kommt.

Diese Arbeit ist auch ein großes Bekenntnis zu freier Software. Freie Software hat Bastelcharakter und ihr haftet das Flaire von verschrobenen Hackern an. Ihr einzigartiger Vorteil ist jedoch ihre Freiheit, dass jeder damit machen darf, was er will, sofern es beim *fair use* bleibt.

Dies steht auch vor dem Hintergrund meiner Überzeugung, dass Information frei sein muss. Und Software ist nichts anderes als ein Werkzeug und Zugang zu und für Information. Daher wird manche Sichtweise in dieser Arbeit auch etwas stur anmuten, aber unter berücksichtigung dieser Philosophie immer noch verständlich.

Stuttgart, im Oktober 2002

Inhaltsverzeichnis

Abbildungsverzeichnis

Tafelverzeichnis

Teil I

Theoretischer Teil

Kapitel 1

Grundlagen

1.1 Warum digitale Bibliotheken?

Digitale Bibliotheken versprechen bei einer bedachten und konsequenten Umsetzung das Beste aus zwei Welten: sie bieten einen qualitativ anspruchsvollen und zuverlässig gepflegten Bestand an Informationsquellen, der durch die leistungsfähige Verarbeitungskapazität moderner Informationstechnik verwaltet und bereitgestellt wird.

Das Angebot an Information weltweit ist inzwischen unüberschaubar geworden. Nicht nur hat die alleinige Masse der zur Verfügung stehenden Quellen zugenommen, auch deren Qualität erstreckt sich über das gesamte Spektrum des Möglichen. Wegweiser in diesem Dschungel an „Exformation“ (in Anlehnung an Lem 1997) und Instrumente um mit dieser Masse umzugehen sind neben der Frage nach der Zuverlässigkeit und Gültigkeit der Inhalte notwendiger als je zuvor.

Überlegungen zu dieser Problematik sind nicht neu (siehe dazu die Beispiele im Abschnitt 1.3 auf Seite 6) und reichen bis in die, aus heutiger Sicht, „Vorzeit“ der Daten- bzw. Informationsverarbeitung zurück. Doch sind digitale Bibliotheken alles andere als rein technische Lösungen.[1] Ihr Bereich liegt in der Schnittstelle aus den klassischen Disziplinen des BID-Bereiches (Bibliothek, Information, Dokumentation) und der Informationstechnik oder Informatik mit einem besonderen Schwerpunkt auf *information retrieval* und *text and data mining*[2]. Zunehmendes Interesse gewinnt dieser Komplex auch durch das umfassende Gebiet der künstlichen Intelligenz und der Idee der Expertensysteme[3].

Die Chance digitaler Bibliotheken liegt in den Synergieeffekten, die sich aus den Kompetenzen der ihnen zu Grunde liegenden Disziplinen ergeben. An die Seite des über Jahrhunderte hinweg entwickelten Erfahrungsschatzes der Bibliothekare und Dokumentare, wie man Information am besten ordnet und zur Verfügung stellt, gesellt sich das Wissen der Informatiker, wie große Datenmengen schnell und effizient verarbeitet werden.

[1] Auch wenn Endres und Fellner (2000) dies in ihrem Buch, wie schon dessen Titel andeutet, anders sehen.

[2] *Information retrieval* und *text and data mining* sind Teildisziplinen der Informatik, die sich mit der maschinellen Verarbeitung von kodierter Information auseinander setzen. Zur Einführung siehe Sparck-Jones (1997)

[3] Expertensysteme sind Computeranwendungen, die durch die Eingabe signifikanter Daten selbstständig Entscheidungen fällen sollen (Haun 2000).

1.2 Begriffe und Definitionen

Digitale Bibliotheken entspringen nicht nur terminologisch den „klassischen Bibliotheken“[4]: Hier wie da liegt ihnen die Idee zu Grunde, Informationsquellen zu sammeln, zu erschließen und ihrer Nutzerschaft bereit zu stellen.

Zwar erinnert der Begriff „Bibliothek“ viele an verstaubte Bücher, düstere Magazine und totenstille Lesesäle, doch erfreut er sich nicht zuletzt wegen seiner prestigeträchtigen Assoziation als „Gedächtnis der Menschheit“ (Jochum 1993, S. 7) großer Beliebtheit. Dementsprechend erfolgt sein Gebrauch vielfach auch bedeutungsfremd und inflationär. Daher sei zunächst ein Blick auf die in der Diskussion um digitale Bibliotheken vorkommenden Begriffe und deren Bedeutung geworfen.

1.2.1 Die Bibliothek und ihre Aufgaben

Eine umfassende und gültige Definition des Begriffes „Bibliothek“ haben Ewert und Umstätter (1997, S. 10 f.) in ihrem Werk geliefert:

> „Die Bibliothek ist eine Einrichtung, die unter archivarischen, ökonomischen und synoptischen Gesichtspunkten publizierte Information für die Benutzer sammelt, ordnet und verfügbar macht.“

Als Kernaufgaben einer Bibliothek zählen demnach *Sammlung*, *Ordnung* und *Bereitstellung*, die unter bestimmten Gesichtspunkten zu erfolgen haben.

Unter **Sammlung** ist der Aufbau eines eigenen Bestandes an Informationsquellen gemeint, der das Herzstück der Bibliothek darstellt. In konventioneller Weise besteht dieser aus Büchern und Zeitschriften, in bedeutendem Maße aber auch aus Mikroformen, AV-Materialien und Datenträgern. Formal gesehen stellt eine Bibliothek prinzipiell eine Dokumentensammlung dar, inhaltlich gesehen eine Daten- oder Informationsbasis[5].

Bibliotheken der dritten Funktionsstufe[6] kommt bereits die Aufgabe der Sicherung bzw. Archivierung zu (BDB 1994, S. 35 ff.). So beziehen sich die o. g. *archivarischen Gesichtspunkte* auch auf die Frage der langfristigen bzw. dauerhaften Archivierung des Bestandes und auf Maßnahmen, diesen auch für die Benutzung verfügbar zu halten.

Das Prinzip der **Ordnung** ist eine klassiche Arbeit von Bibliotheken und Archiven. Dahinter steht das Bestreben, jedes einzelne Objekt des Bestandes zumindest wieder aufzufinden, wie auch eine Kenntnis über dessen Größe und Zusammensetzung zu haben. Der eigentliche Zweck der Ordnung ist jedoch das *gezielte Wiederauffinden* von Objekten im Sinne des *information retrieval* (Gaus 2000, S. 1). Dies kann zum einen ein bestimmtes, bekanntes Objekt sein („Das Buch mit der Signatur abc/...“), aber auch eine unbestimmte Anzahl unbekannter Objekte, die aber alle eine bestimmte Eigenschaft besitzen („Alle Zeitschriften des Jahres 1999“; „Alle Medien zum Thema *'text mining'*“).

Die Tätigkeiten *Sammeln* und *Ordnen* wie auch das *Recherchieren* werden nach Gaus (2000, S. 3) durch den Begriff **Dokumentation** umfasst. An anderer Stelle (Ewert und Umstätter 1997, S. 12) wird Dokumentation jedoch definiert als die

[4] Bibliotheken im herkömmlichen Sinne werden fortan als „konventionelle Bibliotheken“ bezeichnet.

[5] Zur Abgrenzung von Daten, Information und Wissen siehe Abschnitt 1.2.6 auf Seite 6

[6] „Spezialisierter Bedarf: Landesbibliotheken, Hochschulbibliotheken, Spezialbibliotheken, Großstadtbibliotheken“ (BDB 1994, S. 11)

> „Sammlung und Speicherung, Klassifikation und Selektion, Verbreitung und Nutzung von Information jedweder Form."

In herkömmlicher Weise unterscheiden sich hierin die Aufgaben von Bibliotheken und Archiven von denen der *Fachinformationszentren*: erstere erschließen vorangig die „großen" Einheiten ihres Bestandes wie Monografien, letztere widmen sich schwerpunktmäßig unselbständigen Werken, Beiträgen, etc. und grauer Literatur.

Die **Bereitstellung** des Bestandes ist der eigentliche Sinn einer Bibliothek: erst durch dessen Nutzung ist er von Wert und stellt so einen Beitrag zur Versorgung der Benutzer mit Information dar.

Die *ökonomischen Gesichtspunkte* beziehen sich auf die Wirtschaftlichkeit der genannten Tätigkeiten und deren Sinnhaftigkeit. Gemäß BDB (1994, S. 11) wird der Aufgabenkatalog einer Bibliothek bestimmt durch „Zielgruppen und ihren Bedarf". So ist vor der Frage des Aufwandes und damit auch der Kosten für die zu erbringenden Leistungen auch der Nutzen für diese Zielgruppen ein entscheidendes Kriterium.

Unter den *synoptischen Gesichtspunkten* ist laut Ewert und Umstätter von den Bibliotheken „eine Synopsis des gesamten Informationsangebotes dieser Welt herzustellen". Dies gilt als maßgebendes Kriterium des Bestandsaufbaus, indem nicht „einfach alles" in den Bestand übernommen wird, sondern nach Kriterien der Zielgruppen und ihrem Bedarf kritisch ausgewählt wird und die verschiedenen möglichen Angebote gegeneinander abgewogen werden. Dies impliziert auch eine *fachliche Ausrichtung* des Bestandes.

Diese Definition grenzt den Typus einer Bibliothek ab gegen den einer **Buchhandlung**, deren Anliegen der *Verkauf* ihres „Bestandes" ist, ferner gegen den eines **Archives**, dessen primäre und vornehmliche Aufgabe die dauerhafte Archivierung von Originaldokumenten ist (vgl. Rehm 1991, unter „Archiv").

Der unscheinbare aber überaus wichtige Punkt der „publizierten Information" weist Bibliotheken die Aufgabe zu, sich auf bereits existierende Informationsquellen zu beschränken. Die Erzeugung von Information sowie deren Verarbeitung fällt demnach nicht in den Aufgabenbereich von Bibliotheken.

1.2.2 Digitale Bibliothek

Unter einer digitalen Bibliothek soll hier das digitale Pendant einer konventionellen Bibliothek verstanden werden, wie dies auch von McKnight (1997), allerdings unter dem Titel *„electronic library"*[7] definiert wird:

> „The concept of information stored electronically and made accessible to users through electronic systems and networks, but having no single physical location. It is, therefore, analogous to a library as a storehouse of information, but has an existence in virtual reality."

Das Prinzip einer konventionellen Bibliothek lässt sich nicht 1 : 1 in das digitale Umfeld übertragen. Gemäß Endres und Fellner (2000, S. 4 f.) beschäftigt sich eine digitale Bibliothek mit Beständen und Objekten in digitaler Form, also mit computerlesbar vorliegenden Dokumenten[8]. Desweiteren sollen sämtliche in Abschnitt 1.2.1 angeführten und für eine konventionelle Bibliothek gültigen Kriterien auch für eine digitale Bibliothek gelten.

[7] Der hier anders verstandene Begriff „elektronische Bibliothek" wird in Abschnitt 1.2.3 auf der nächsten Seite behandelt.

[8] Für eine Bestandseinheit wird nachfolgend der Begriff *Dokument* verwendet.

Da der Entwurf eines modellhaften Konzeptes Gegenstand dieser Arbeit ist, werden die näheren Kriterien hierzu in Kapitel 2 auf Seite 12 erarbeitet.

1.2.3 Elektronische Bibliothek

Eine *elektronische Bibliothek* beschreibt eine Bibliothek, deren Arbeitsgänge mit Hilfe elektronischer Maßnahmen erledigt oder unterstützt werden. Gemeint sind hier vor allem automatisierte Geschäftsgänge, datenbankgestützte Katalogisierung und vergleichbare Tätigkeiten sowie das Angebot sogenannter *„neuer Medien“* im Bestand.[9] Somit können mittlerweile Bibliotheken an sich, mit einigen Ausnahmen allerdings, durchweg als „elektronische Bibliotheken“ bezeichnet werden. In diesem Fall hat die allgemeine Entwicklung eine besondere Bezeichnung und somit den Begriff als solchen überflüssig gemacht.

1.2.4 Virtuelle Bibliothek

Eine *virtuellen Bibliothek* ist als Entität nicht existent. Prominentestes Beispiel ist die *World Wide Web Virtual Library*[10], die vom WWW-Erfinder Tim Berners-Lee initiiert wurde. Eine virtuelle Bibliothek ist dadurch charakterisiert, dass sie selbst keinen Bestand besitzt und sich auch durch keinen Ort bestimmen lässt[11]. Zwar haben auch Netzwerkdienste einen Standort, doch liegen hier die Inhalte, um die es eigentlich geht, physikalisch strukturlos verteilt, im angeführten Beispiel sogar rund um die Welt.

Hintergrund einer virtuellen Bibliothek ist der Gedanke, verstreut vorliegende Informationsquellen unter „einem Dach“ zusammen zu führen. Dem Benutzer steht so eine stimmige Oberfläche und Schnittstelle zur Verfügung, während die eigentlichen Inhalte verstreut auf anderen Servern liegen.

Eine virtuelle Bibliothek kann daher auch konventioneller Natur sein: denkbar wäre ein einheitlicher OPAC für die Bibliotheken eines Ortes oder einer Region, was durch die Verbundkataloge bereits umgesetzt ist. Dem Benutzer erscheinen die Bestände mehrerer Bibliotheken konsistent unter einem Zugang, wie es für den Bestand einer einzelnen Bibliothek auch der Fall wäre.

1.2.5 Hybride Bibliothek

Spätestens seit dem Gutachten des Wissenschaftsrates (Wissenschaftsrat 2001) ist der Begriff „hybride Bibliothek“ recht populär geworden. Dahinter verbirgt sich nichts anderes als eine konventionelle Bibliothek, die ihr Angebot um das einer digitalen Bibliothek erweitert. In Abschnitt 2.8 auf Seite 29 wird näher auf den Aspekt der Trägerschaft eingegangen, vorerst sei daher gesagt, dass sich dem hier vertretenen Verständnis nach eine digitale Bibliothek nicht unabhängig von einer sie unterstützenden oder tragenden konventionellen Bibliothek oder vergleichbaren Einrichtung sehen lässt, wie sich auch konventionelle Bibliotheken seit langer Zeit mit elektronischen Publikationen beschäftigen. Dieser Begriff wird hier künftig nicht verwendet.

[9] nach Rehm (1991, unter „elektronische Bibliothek“)

[10] `http://www.vlib.org/`

[11] Dies kann auch eine Eigenschaft digitaler Bibliotheken sein.

1.2.6 Daten – Information – Wissen

Im Bereich der digitalen Bibliotheken tauchen immer wieder Begriffe wie „Datenverarbeitung“, „Informationsvermittlung“ oder „Wissensmanagement“ auf. Um Klarheit über die Ebenen zu erhalten, die hier mit hineinspielen, seien die Begriffe *Daten*, *Information*[12] und *Wissen* gemäß der Definition von Gundry (2001) gegeneinander abgegrenzt:

Daten sind simple Repräsentationen von Fakten wie Uhrzeiten, Temperaturen, Wasserständen, etc.

Information besteht aus Daten, die in einem Zusammenhang stehen und für ein einzelnes Individuum relevant sind, z. B. Abfahrtszeiten

Wissen umfasst schließlich sämtliche Einsichten und Fähigkeiten eines Individuums zu Handeln, die sich aus dem Erwerb von Information ergeben.

Da Information demnach immer nur aus der Interpretation von Daten besteht, könnte an sich vom Bestand einer Bibliothek[13] nicht als „Information“ die Rede sein, sondern müsste sich stets auf Daten beziehen.

Da das Grundprinzip einer Bibliothek aber die Nutzung ist, entwickeln sich alleine hieraus die im Bestand verfügbaren Daten zu Information. In diesem Sinne kann durchaus von einem Bestand an Information gesprochen werden.

1.3 „Wissen aus der Maschine“

Die prinzipielle Idee einer digitalen Bibliothek ist die maschinelle Verwaltung von Informationsressourcen oder Information selbst, die wesentlich älter ist als sämtliche modernen und populären Ansätze.

Im Folgenden soll ein Blick auf die prominentesten Ideen geworfen werden, um die Vorstellung der Probleme und Ansätze zu deren Lösung darzustellen.

1.3.1 H. G. Wells: World Brain

1937 entwarf der bekannte Science-Fiction-Autor H. G. Wells die Idee einer permanenten, globalen Enzyklopädie, die dezentral von Wissenschaftlern, Forschungseinrichtungen und Universitäten gepflegt wird (Wells 1938). Auf diese Weise würde eine von einer zentralen Instanz weitgehend unabhängige Wissensbasis geschaffen werden, die von der Idee her all jenes Wissen enthält, über das die beitragenden Personen und Institutionen verfügen. Dieser Bestand sollte sämtlichen Menschen offen stehen. Allen voran maß Wells Wissenschaftlern, Politikern und Journalisten die größte Nutzung zu.

Wells' Ausgangspunkt waren die technologischen Entwicklungen der Radio-, Fernseh- und Fototechnik, doch beschränkt er sich beim Entwurf seines Modells auf die konzeptionellen Überlegungen, ohne in weiterer Weise auf dessen Umsetzung oder die zu verwendenden Techniken einzugehen.

So steht hinter Wells' Idee diejenige einer weltweiten, verteilt gepflegten Informationsquelle, die in gewisser Weise durch das Internet greifbar geworden ist. Zwar wurde Wells'

[12] Von Information ist hier als anonyme Entität die Rede, daher stets auch in Singularform.

[13] Sofern nicht differenziert, umfasst „Bibliothek“ sowohl konventionelle als auch digitale.

Konzept vor einiger Zeit noch akademisch verfolgt (etwa bei Goodman 1987), spielt ansonsten jedoch eine untergeordnete Rolle. Als Umsetzung von Wells' Idee können beispielhaft das ursprünglich „GNUPedia“ betitelte *Free Encyclopedia Project*[14] und die dort genannten Online-Enzyklopädien WIKIPedia[15] und NUPedia[16] gelten.

1.3.2 Vannevar Bush: Memex

Mit *Memex* (für „Mem*ory* Ex*tension*) entwarf Vannevar Bush die Vision von 10 000 Seiten der *Encyclopædia Britannica*, die auf einer DIN A 4 Seite gespeichert werden können (Bush 1945). Damit sollte dem Benutzer der praktisch ortsungebundene Zugriff auf diese Informationsquellen gestattet werden. Gemäß den Mitteln seiner Zeit bezog Bush dies noch auf Mikrofilm-Technik.

Grundlage für Bushs Ideen war die Arbeitsweise des menschlichen Gehirns, das von einer Idee aus zu weiteren springt, und so zwischen zwei Punkten Assoziationen und Beziehungen bildet. Die Nachbildung dieser Beziehungen im Datenbestand einer Enzyklopädie sollte durch Memex möglich sein.

Diese Verknüpfungen, *„trails“* genannt, könnten auch durch einen neuen Berufsstand, dem des *trail blazers*, erzeugt werden. Auch dachte Bush nicht an reine mikroverfilmte Werke sondern an eine geordnete und klassifizierte Sammlung, deren Pflege ebenfalls Aufgabe dieses Berufsstandes wäre.

Memex ist somit zweischichtig zu sehen: Die Grundlage bildet der Datenbestand, der so gesehen „atomarer“, also grundlegender und nachbildbarer, Natur ist. Die zweite Komponente wären die Verknüpfungsnetze, die sich aus der Arbeit mit einem Memex-System ergeben. In diesen finden sich stark abstrahiert die Gedankengänge des Benutzers wieder und sind dementsprechend einzigartig. Der grundlegende Datenbestand wird somit quasi um das *„Wissen“* des Nutzers angereichert.

Das Prinzip von Objekten und Verknüpfungen verfolgte auch *Enquire*, das erste von Tim Berners-Lee geschriebene hypertextbasierte Programm, das als Vorläufer zu seiner Web-Idee gelten kann.[17]

1.3.3 Douglas Engelbart: Augment

„Augment“ von Douglas Engelbart verstand sich als Methode zur Verbesserung der geistigen Leistungsfähigkeit eines Menschen, indem seine Gedankengänge auf Lochkarten nachgebildet und diese Strukturen dauerhaft gespeichert werden (Engelbart 1962).

Im Gegensatz zu Bush ging Engelbart bei seinem Entwurf auf die Leistungsfähigkeit der maschinellen Datenverarbeitung ein und richtete sein Projekt auch daraufhin aus. So sah er einen überragenden Vorteil des von ihm beschriebenen Systems in der Schnelligkeit, mit der Assoziationen gebildet und verfolgt werden können.

Im späterhin aufgekommenen Prinzip der *Mikrofiche-Lochkarten* (siehe Abbildung 1.1), das Mikrofilm-Aufnahmen über Lochkarten-Indizes maschinenlesbar erschloss, lässt sich Engelbarts Idee zwar noch erahnen, ging ansonsten aber mit dem Lauf der technischen Entwicklung unter. Während seiner weiteren Beschäftigung mit den Prinzipien des Hypertext, dessen

[14] `http://www.gnu.org/encyclopedia/index.html`

[15] `http://www.wikipedia.com/`

[16] `http://www.nupedia.com/`

[17] Eine Beschreibung dessen findet sich bei Berners-Lee und Fischetti (1999, Kap. 1). Das damalige Handbuch zu *Enquire* ist als Scan unter `http://www.w3.org/History/1980/Enquire/` zu finden.

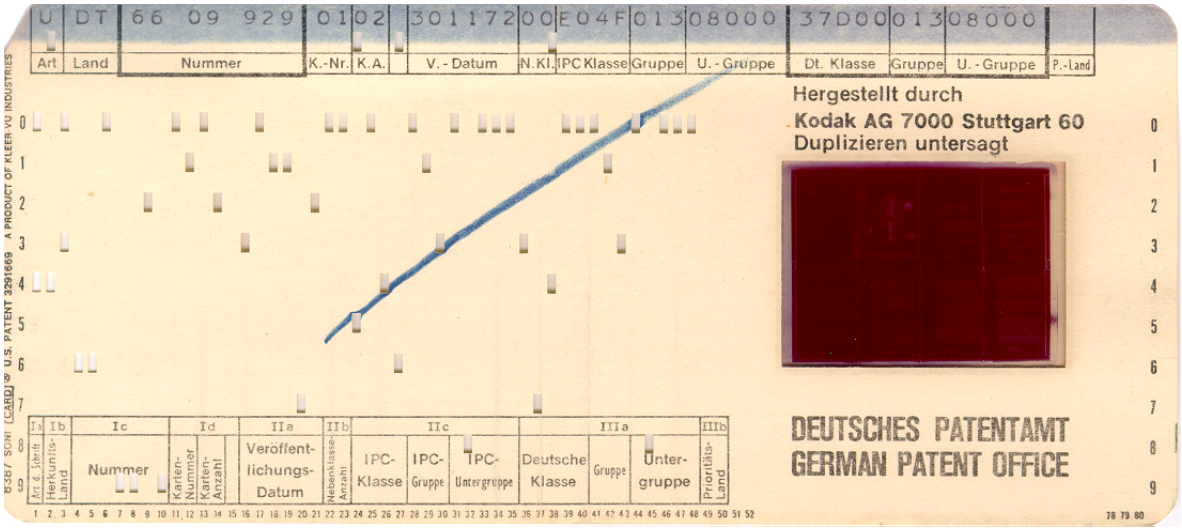

Abbildung 1.1: Mikrofiche-Lochkarte des Deutschen Patentamtes

Begriff erst später durch Ted Nelson geprägt wurde, ging Engelbart als der Erfinder der Computer-Maus in die Geschichte ein.

1.3.4 Ted Nelson: Xanadu

Ted Nelsons visionäres Projekt vom nicht-linearen Text (Nelson 1992) hat zwar zahlreiche weitere Projekte beeinflusst[18], es selbst aber zu keiner Produktionsreife gebracht[19], was Anlass zu einem oftmals als bösartig bezeichneten Artikel in der Zeitschrift *Wired* war (Wolf 1995).

Kernstück von *Xanadu* ist der nicht-lineare Text, für den Nelson den Begriff *Hypertext* und für nicht-textuelles Material den der *Hypermedia* prägte. Genutzt werden sollte diese Möglichkeit, um aus verschiedenen Textbausteinen neue Texte zu erstellen oder Textstücke zu zitieren, ohne diese kopieren zu müssen. Für die Nutzung dieses Konzeptes dachte Nelson an geringe Zitiergebühren. Die Ideen des *transcopyright* und *transquoting* beruhen hierauf.

Auch wenn die seit geraumer Zeit wieder aufgenommenen Versuche einer Umsetzung dieser Ideen, diesmal auf Grundlage des WWW, ebenfalls noch keine Prototypen hervorbringen konnten, ist durch die Möglichkeiten der an XML gebundenen XLink und XPointer deren Realisation bereits möglich.[20]

1.3.5 Tim Berners-Lee: World Wide Web

Tim Bernes-Lee entwickelte das World Wide Web als Instrument, den Wissenschaftlern des *Conseil Européen pour la Recherche Nucléaire* (CERN)[21] die Veröffentlichung und Verteilung ihrer Dokumente zu erleichtern. Hintergrund und Antrieb war die Idee, jedem Teilnehmer die

[18] Siehe `http://xanadu.com.au/projects.html`

[19] Einzelne Programme, die diese Ideen teilweise oder stückhaft umsetzen finden sich unter `http://www.xanadu.net` und `http://www.udanax.com`.

[20] Siehe hierzu die XML-Seiten des W3C ⟨`http://www.w3.org/XML/`⟩.

[21] `http://www.cern.ch/`

unkomplizierte Bereitstellung seiner Dokumente zu ermöglichen (Berners-Lee und Fischetti 1999, Kap. 2).

Daher hängen der Entwicklung des WWW auch weniger inhaltliche Überlegungen als technische an. Die Resultate Berners-Lees Arbeit sind im wesentlichen die Auszeichnungssprache HTML, die das Prinzip des Hypertext realisiert, und das Kommunikations-Protokoll HTTP.

Das Ergebnis kann durchaus als Revolution des Datenverkehrs bezeichnet werden. Der weltweite Einsatz dieser Techniken führte zu einem unermesslichen Volumen an Daten, das sich im Grunde ohne kontrollierende, regelnde oder ordnende Instanz im Internet, genauer gesagt im WWW, findet.

Der Erfinder des Web selbst gibt zu, dass dieses „dumm“ sei und will mit dem Ansatz des *„Semantic Web“* Intelligenz in die Hypertext-Strukturen bringen.[22]

1.3.6 Die tragenden Ideen

Als tragende Ideen, gleich ob diese verwirklicht wurden oder nicht über das gedankliche Reißbrett hinaus kamen, können aus diesen Konzepten festgehalten werden:

- schneller und einfacher Zugriff auf Informationsquellen von praktisch jedem beliebigen Platz aus
- dezentrale und weltweit verteilte Speicherung der Daten
- Verknüpfung der gespeicherten Informationen zu sogenannten *trails* (Informationspfaden)
- Lösung des Textes von seiner Linearität
- problemlose und ungebundene Zitierung
- nicht regulierte Publikationsmöglichkeit
- Ordnung und Klassifizierung der Inhalte

Bereits realisiert oder in Kürze flächendeckend[23] zu realisieren sind vor allem die technischen Komponenten. Bedarf zur Konzeption wie auch zur Umsetzung besteht vor allem in den inhaltlichen Aspekten.

1.4 Probleme und Chancen

Das *Web Characterization Project*[24] von OCLC stellt zuverlässige Daten über das Web und seine Inhalte regelmäßig zur Verfügung und ermöglicht so einen guten Überblick über dessen Größe und Zusammensetzung. Tafel 1.1 zeigt das jährliche Wachstum an Web-Sites wie auch deren Gesamtzahl.[25]

[22] Siehe dazu Berners-Lee u. a. (2001) sowie die *„Semantic Web Activity“* des W3C ⟨`http://www.w3.org/2001/sw/`⟩ und das *„Semantic Web Community Portal“* ⟨`http://www.semanticweb.org/`⟩.

[23] Funktionierende Laborlösungen sind zwar nicht selten, zur Akzeptanz von Techniken und Verfahren müssen diese aber allen zugänglich sein.

[24] `http://wcp.oclc.org/`

[25] Die Daten entsprechen den am 2002-09-07 verfügbaren.

Zeitraum	Wachstum	Anzahl
1997 – 1998	82 %	2 851 000
1998 – 1999	71 %	4 882 000
1999 – 2000	52 %	7 399 000
2000 – 2001	18 %	8 745 000
1997 – 2001	457 %	– / –

Tafel 1.1: Jährliches Wachstum der Anzahl an Web-Sites

Anteil	Inhalte
36,3 %	akademisch
63,7 %	nicht-akademisch

Tafel 1.2: Anteil akademisch eingestufter Web-Sites

Allein schon das 2001 verfügbare Angebot von 8 745 000 Web-Sites schlägt mit einer immensen Anzahl zu Buche. Da es sich hier jedoch um Gesamtangebote handelt, die durchaus aus 100 oder mehr Einzelseiten bestehen können, wird das Problem der Masse sehr schnell deutlich. So bietet der Internet-Suchdienst *Google*[26] im September 2002 nach eigenen Angaben auf der Startseite das „Suchen auf 2 469 940 685 Web-Seiten“ an, worin aber auch Usenet-Archive enthalten sind.

Wenn man diese Zahl von rund 2,5 Milliarden Seiten mit der Klassifikation der Inhalte vergleicht, wie sie OCLC ebenfalls anbietet, nach der nur 36,3 % der Inhalte akademischer Natur sind (siehe Tafel 1.2)[27], so können aus dieser Sicht etwas weniger als zwei Drittel des Web-Inhaltes für die wissenschaftliche Arbeit als nicht relevant gelten.

Dies bedeutet immerhin ein Volumen von (einfältig gerechnet) rund 1,6 Milliarden Seiten, die nicht relevant aber zugänglich sind und anteilsmäßig bei einer Recherche in den einschlägigen Suchdiensten auftauchen.

Dass einfache Suchstrategien, mangelnde Medienkompetenz und die daraus resultierende Unfähigkeit, relevante Inhalte von nicht-relevanten zu unterscheiden und Ergebnisse bewerten zu können ein nicht geringes Problem darstellt, wurde zuletzt in der sogenannten „Stefi-Studie“ (Klatt u. a. 2001) wieder verdeutlicht.

Diesem Problem können digitale Bibliotheken mit ihrem Prinzip der Auswahl, Bewertung und Erschließung begegnen. Die Ungenutztheit dieser Chancen erkannte auch der Wissenschaftsrat in seinem Gutachten zur digitalen Informationsversorgung (Wissenschaftsrat 2001) und sieht gerade hier ein hohes Entwicklungspotential hinsichtlich der Unterstützung und Versorgung mit elektronischer (Fach-)Information.

Dem Problem der Masse ist in technischer Hinsicht durch digitale Bibliotheken einfach zu begegnen. Die mögliche Effizienz maschineller Datenverarbeitung hat die Perspektiven der Erfinder der ersten Stunde wie Bush oder Engelbart inzwischen bei weitem übertroffen.[28]

[26] `http://www.google.com/` sowie `http://www.google.de/`

[27] Als „akademisch“ gelten die Gruppen „Information“, „Professional, Scientific and Technical Services“ und „Educational Services“ aus der Original-Statistik.

[28] Ein Beispiel zum Fortschritt der Technik: 1996 war ein Notebook mit einem 486DX/4-Prozessor mit 75 MHz Taktung, 8 MB Arbeitsspeicher und 520 MB Festplattenkapazität Mittelklasse. 2002, 6 Jahre später, belaufen

Insofern sind die Grundlagen bereits gelegt. Doch wie auch Klatt u. a. und der Wissenschaftsrat erkennen, besteht an inhaltlichen Gestaltungen und Leistungen ein erheblicher Mangel und dringender Nachholbedarf.

sich diese Daten auf 1,8 GHz Taktung, 256 MB RAM und 40 GB Festplattenkapazität.

Kapitel 2

Konzeption

Der kritische Faktor bei der Umsetzung digitaler Bibliotheken ist weniger die Technik. Vielmehr müssen die durch sie eröffneten Möglichkeiten auch sinnvoll genutzt werden, was angesichts der Fülle und Reichhaltigkeit an Möglichkeiten oftmals an der Unübersichtlichkeit scheitert.

Dem entgegen wirken strategische und konzeptionelle Überlegungen, die ein einerseits starres und definitives Raster der umzusetzenden Eigenschaften, andererseits aber auch einen flexiblen Richtungsweiser an die Hand geben. Eine Konzeption definiert so klare Ziele, denen die technische Umsetzung zu genügen hat.

Ziel dieses Kapitels ist es, einen Satz von Kriterien zu formulieren, die bei der Implementierung einer digitalen Bibliothek als Pflichtenheft gelten sollen.

Zwar wahrgenommen, aufgrund der hohen Spezifität bei konkreten Implementierungen hier aber nicht ausgeführt, sind die Aspekte **Personal**, **Organisation** und **Finanzen**.

2.1 Aufgaben

Wie angemerkt sollen die Merkmale einer konventionellen Bibliothek (vergleiche Abschnitt 1.2.1) auch gleichzeitig Kriterien für die hier zu konzipierende digitale Bibliothek sein.

Hacker (1992, S. 20) äußert sich zu den Aufgaben einer Bibliothek wie folgt:

> „Als Literatur-, Bücher- und Mediensammlungen, als spezielle Informationseinrichtungen im Dienstleistungsbereich haben die Bibliotheken den Auftrag, ihren Benutzern Literatur, Bücher und Medien kostenlos und leihweise zur Verfügung zu stellen. *Ihre Hauptaufgaben bestehen in der Literaturversorgung und der Literaturinformation,* zusätzliche Aufgaben ergeben sich durch die Vermittlung von Medien und Informationen über den Literaturbereich hinaus (Musik, Bilder, Filme).“

Die primäre Aufgabe ist auch hier die Literaturversorgung und -information. So widmet sich auch eine digitale Bibliothek primär der Literatur- bzw. der Informationsversorgung[1], gemäß ihrer Natur betrifft dies digitales Material. Den Einheiten „Literatur, Bücher und Medien“ sei hier daher durch die globale Bestandseinheit „Dokument“ entsprochen.

Während sich das Prinzip „leihweise“ durch die beliebige und unbegrenzte, wie vor allem ohne Qualitätsverlust mögliche Duplizierbarkeit digitalen Materials erübrigt, muss dieser

[1] Zur Diskussion um Literatur- und Informationsversorgung siehe Abschnitt 2.6 auf Seite 26.

Punkt hinsichtlich des rechtlichen Rahmens und der neuen Möglichkeiten eingehender betrachtet werden.[2]

Kriterium 1 *Eine digitale Bibliothek stellt die Versorgung ihrer Klientel mit digital vorliegender Literatur und Information sicher.*

Mit dem Begriff „Klientel" seien hier zunächst die Benutzer allgemein gemeint. Zur näheren Abgrenzung dieses Begriffes sei auf Abschnitt 2.9 auf Seite 30 verwiesen.

Hacker nennt den Aspekt „kostenlos", der für diese Konzeption eine besondere Gewichtung bekommen soll. Der in Artikel 5 des Grundgesetzes der Bundesrepublik Deutschland garantierte Grundsatz der freien Information setzt „allgemein zugängliche Quellen" (Deutschland 1949, Art. 5) voraus. Diese allgemein zugänglichen Quellen stellen in erster Linie Bibliotheken mit ihrem aktuellen, breitgefächerten und vor allem frei und kostenlos zugänglichen Bestand dar (vgl. BDB 1994).

So besteht die Idee einer allgemein zugänglichen Bibliothek darin, einer breiten Öffentlichkeit den Zugang zu seltener, teurer und schwer beschaffbarer Literatur zu ermöglichen. Angesichts der fortschreitenden Kommerzialisierung der im Internet angebotenen Inhalte und Dienste muss dieses Prinzip der freien Verfügbarkeit auch für diesen Bereich als unabdingbar erachtet werden. In diesem Aspekt stellt der freie Zugang zu digitalen Ressourcen die logische und konsequente Weiterentwicklung der Praxis und des Dienstleistungsangebotes einer Bibliothek dar. Jedoch bedarf dieser Punkt einer besonderen Fragestellung hinsichtlich der rechtlichen Aspekte (vgl. Abschnitt 2.10 auf Seite 30).

Kriterium 2 *Eine digitale Bibliothek stellt ihren Bestand und ihre Dienstleistungen im Rahmen der Möglichkeiten ihrer Klientel kostenlos zur Verfügung.*

2.2 Archivierung

Die Archivaufgabe großer Bibliotheken hat ihren Grund: sie stellen so die – nach Möglichkeit – dauerhafte Verfügbarkeit von Information sicher. Besonders im wissenschaftlichen Bereich kommt dieser Aufgabe der unverzichtbare Wert der Quellensicherung zu. In Disziplinen, die von Zitaten leben, bzw. in denen alte Erkenntnisse immer noch von Bedeutung sind, ist die Wiederauffindbarkeit der Originale von größter Bedeutung. Eine dauerhafte und zuverlässige Verfügbarkeit garantiert eine korrekte Zitierbarkeit und ein problemloses Heranziehen der Quelle.

Bei digitalen Materialien ist durch die Archivierung deren Fortbestand in vielen Fällen überhaupt erst gesichert. Angesichts der niedrigen Halbwertszeit von Internet-Adressen (vgl. Meyer 2002) ist es gerade für wertvolle Ressourcen von Bedeutung, diese dauerhaft und zuverlässig zu archivieren und möglichst auf Dauer unter einer Adresse zu speichern.

In seinem Sinne versucht das *Internet Archive*[3] dies mit Web-Inhalten zu tun: zu bestimmten Zeiten werden Momentabzüge, sogenannte *„snapshots"* von Servern erstellt und archviert. Die inzwischen fast schon klassischen „404"-Fehlermeldungen der Webserver („Page not found") sind damit zwar nicht lösbar, doch will das Projekt eine Möglichkeit geben, wenigstens noch *eine* Version von den entsprechenden Dateien aufzufinden.

[2] Siehe dazu Abschnitt 2.10 auf Seite 30.

[3] `http://www.archive.org/`

Dieses Modell weist aber aus bibliothekarischer Sicht einen Mangel an Dokumentation und Systematik auf. So ist der Datenbestand ein rein gespiegeltes Abbild des Bestandes im Web.

Kriterium 3 *Eine digitale Bibliothek stellt die langfristige und möglichst dauerhafte Archivierung ihres Bestandes sowie dessen technische Zugänglichkeit sicher.*

Diese Aufgabe erfordert jedoch weit mehr als die Bereitstellung von technischer Infrastruktur. Dem technischen Wandel muss nicht nur im Bereich der Hardware Rechnung getragen werden, es muss auch sichergestellt werden, dass die archivierten Dokumente auf Dauer mit aktueller Technik, vorrangig geht es hier um Software, zu verarbeiten sind.

Nach Henze (1999) bestehen drei Möglichkeiten, um die dauerhafte Verfügbarkeit digitaler Information sicherzustellen:

Präservierung beinhaltet neben der Archivierung der Dokumente selbst auch die der Originalumgebung: Computersysteme und die benötigte Software werden ebenfalls aufbewahrt, um die Verarbeitung der entsprechenden Dokumente mit einem Originalsystem zu ermöglichen. Als problematisch erweist sich hierbei die daraus resultierende „Inselexistenz“ der Materialien. Ihre Nutzung wird im schlimmsten Fall nur noch vor Ort möglich sein[4] und dies nur so lange, als die archivierten Systeme funktionstauglich sind und auch Personal vorhanden ist, das diese Technik beherrscht. Da dies auch nicht dauerhaft der Fall sein wird, zögert diese Methode den Punkt der Unzugänglichkeit der Materialien nur hinaus. Ein Beispiel hierfür sind die noch nicht allzu alten 5,25-Zoll-Disketten, deren Technik zwar bekannt und eigentlich leicht zu fertigen ist, die jedoch von der Entwicklung überholt wurden und außer in kleinen „Restbeständen“ nicht mehr zu finden sind, wodurch Daten auf diesen Datenträgern in dem Sinne schon nicht mehr nutzbar sind.

Emulation versucht, die notwendige Hard- und Softwareumgebung auf neuen Systemen nachzustellen. Dazu müssten zusammen mit den Dokumenten auch Informationen über Geräte und Software archiviert werden. Ein Beispiel sind Computerprogramme, die z. B. auf GNU/Linux-Rechnern die Umgebungen alter Amiga-Systeme nachbilden.[5] Damit sind Programme für diese Plattform zwar benutzbar, jedoch müssen diese zuerst in einer für den PC und den Emulator lesbaren Form vorliegen[6]. Daher sind auch dieser Methode Grenzen gesetzt.

Migration oder „Umkopie“ umfasst das Übertragen der Dokumente von einer Plattform auf eine andere. Dies kann im einfachsten Fall ein Umkopieren zwischen Datenträgern und über Systemgrenzen hinweg sein, kann aber auch eine komplette Konvertierung von einem Dokumentformat zu einem anderen bedeuten. In solch einem Fall bedeutet Migration eine Veränderung des Materials. Dabei wird das Originaldokument zwar nicht in seiner ursprünglichen Form erhalten, wohl sind die Inhalte aber noch verfügbar, teils jedoch mit Einschränkungen. Auch wenn gegenüber dem Original hierbei Abstriche gemacht werden müssen, so stellt diese Methode doch das Vorgehen dar, das dem Zweck der Archivierung, nämlich die dauerhafte *Verfügbarkeit* und *Benutzbarkeit*, weitgehend gerecht wird.

[4] Dies stellt dann auch keine Wertsteigerung gegenüber konventionellen Bibliotheken dar. Vgl. dazu auch Abschnitt 2.7 auf Seite 27.

[5] Siehe dazu `http://www.freiburg.linux.de/~uae/`

[6] Das Betriebssystem GNU/Linux etwa kann Inhalte von Amiga-Festplatten lesen, jedoch nicht jene von Disketten.

Um die Verfüg-, Benutz- und Lesbarkeit von Dokumenten durch Migration sicherzustellen, wie auch den technischen, rechtlichen und finanziellen Aufwand so gering als möglich zu halten, ist bei der Speicherung der Dokumente auf geeignete Speicherformate zu achten.

Kriterium 4 *Bei der Speicherung von Dokumenten sind offene und standardisierte Formate einzusetzen, die eine Skalierbarkeit ermöglichen.*

Eine kurze Übersicht über die Möglichkeiten verschiedener Dokumentformate sowie einigen weiteren Kriterien hierzu gibt Anhang A auf Seite 60.

2.3 Bestand

Eine der bedeutendsten Fragen ist diejenige nach dem Bestand. Während konventionelle Bibliotheken in der Regel historisch gewachsene Bestände haben, können digitale Bibliotheken auf keine solche Grundlage zurückgreifen, sondern müssen häufig bei Null anfangen.

Die Möglichkeiten des Bestandsaufbaus für eine digitale Bibliothek beschränken sich ihrer Natur gemäß auf digital vorliegendes Material. Grundsätzlich ergeben sich hierbei drei Möglichkeiten:

1. der Betrieb als **Publikationsserver** für die Bibliothek oder die tragende bzw. angeschlossene Institution
2. die Sammlung externer Dokumente nach der Art **klassischer Bibliotheksarbeit**
3. die **Digitalisierung** konventionellen Materials

2.3.1 Publikationsserver

Die Arbeit als Publikationsserver ist die einfachste Methode, um einen Bestand aufzubauen und zu erweitern. Dies kann gebunden an eine Institution oder in vollkommen freier Weise erfolgen.

Beispiele für gebundene Publikationsserver finden sich häufig und meist auch unter entsprechenden Bezeichnungen wie „digitale Bibliothek“. Als Publikationsserver enthält ein solches System im Idealfall alle digital vorliegenden Dokumente der betreffenden Institution. Am Beispiel der elektronischen Hochschulschriften[7] wären dies die Veröffentlichungen der Hochschulen, Fachbereiche, Institute als auch der Diplomanden, Doktoranden, etc.

Als Publikationsserver dienen digitale Bibliotheken so der Veröffentlichung, Verteilung und Archivierung der eigenen Publikationen oder zumeist denen der angeschlossenen oder tragenden Institutionen. Digitale Bibliotheken übernehmen damit quasi die Aufgabe eines *Verlages*.

An eine qualitative und inhaltlich kontrollierte Entwicklung des Bestandes ist dabei weniger zu denken, da sich die digitale Bibliothek im Voraus bereit erklärt und verpflichtet, sämtliche eingehenden Dokumente zu veröffentlichen. Eine Kontrolle und Überprüfung des Bestandes, wie ihn die Fachreferatsarbeit bietet, ist hier nicht möglich. Zudem ist der Informationsgehalt für die *internen Nutzer*, also die der angeschlossenen oder tragenden Institutionen, geringer, da der Bestand einzig aus der eigenen Produktion besteht und somit der grundlegende Bedarf an Information nicht gedeckt wird.

[7] Siehe `http://deposit.ddb.de/netzpub/web_online-hochschulschriften.htm` und Maile und Scholze (1997)

Als freier Publikationsserver seien *The Los Alamos E-print Archives*[8] angeführt, die erstmals 1991 von Paul Ginsparg am Los Alamos National Laboratory eingerichtet wurden. Der Dienst fungiert als Preprint-Server für Publikationen der Bereiche Physik, Mathematik und verwandter Wissenschaften und steht Nutzern weltweit offen. 1996 griffen 35 000 Nutzer aus 70 Ländern 70 000 Mal täglich auf das Archiv zu (nach Arms 2000, S. 28).

Kriterium 5 *Eine digitale Bibliothek hat ihre Möglichkeit als Publikationsserver zu fungieren zu nutzen. Allerdings ist hierbei auf die inhaltliche Entwicklung des Bestandes zu achten.*

2.3.2 Externes Material

Die Sammlung von externem Material wird bislang wenig verfolgt, weil sie durch die verschiedenen zu beachtenden Schutzrechtsarten doch einen erheblichen Aufwand bei der Bearbeitung von Material als auch ein gewisses Risiko bei dessen Bereitstellung bedeutet, doch entspricht sie in übertragenem Sinne der Tätigkeit konventioneller Bibliotheken. Anstelle von Büchern, Zeitschriften und weiteren Materialien werden digitale Dokumente gesammelt, erschlossen und archiviert.

Ziel dieser Strategie ist es, aus den verstreut vorliegenden und immer wieder kurzlebigen Materialien aus Internetquellen einen lokalen, geordneten und dauerhaften Bestand aufzubauen. Die digitale Bibliothek wird damit der Aufgabe der Archivierung und Erschließung gerecht.

Ein naheliegendes Beispiel ist die vorliegende Arbeit: sinnvoll wäre es, zusätzlich zu ihrer digitalen Fassung auch die zitierten Online-Ressourcen zu archivieren und die Beziehungen zu verdeutlichen (also *„trail blazing“* zu betreiben).

Das Sorgenkind bei der Beschaffung von Dokumenten sind weniger jene Materialien, die durch Datenbanken oder Volltextserver bereits in *irgendeiner* Weise erschlossen sind, sondern vielmehr der hohe Anteil an so zu bezeichnender *„grauer Literatur“*. Dies betrifft vor allem Dokumente auf den Servern von Wissenschaftler oder Experten, einzelne Veröffentlichungen auf Web-Seiten von Fakultäten und dergleichen, die ansonsten nur über Web-Suchdienste auffindbar wären.

Hier gilt es, dieses relevante Material zu finden, zu erschließen und zu archivieren. Dieses Prinzip beschreibt Pitschmann (2001) ausführlich.

Auf diese Weise leistet eine digitale Bibliothek einen äußerst wertvollen Beitrag zur Informationskommunikation, da bestehenden Publikationen eine größere Verbreitung ermöglicht wird und ansonsten eher marginal wahrgenommene Publikationen sachgerecht erschlossen und in einen fachlichen Kontext gestellt werden.

Kriterium 6 *Eine digitale Bibliothek sammelt externes digitales Material nach inhaltlichen Gesichtspunkten, erschließt dieses und archiviert es in ihrem Bestand.*

Die Integration von externem Material bedingt nicht nur eine verstärkte Auseinandersetzung mit geltendem Recht (siehe dazu Abschnitt 2.10 auf Seite 30), sondern auch mit der Frage der Qualität der Publikationen.

Die Entscheidung über die Aufnahme eines Dokumentes in den Bestand muss einheitlich und nachvollziehbar geschehen. Im Falle eines Publikationsservers ist dies durch den Grundsatz „alles, was aus dieser Institution stammt“ definiert. Bei externen Dokumenten sind diese

[8] `http://www.arxiv.org/`, deutscher Spiegel unter `http://de.arxiv.org/`

Entscheidungen jedoch inhaltlich bezogen. So dürfen die Maßstäbe hierfür nicht gefühlsmäßig gesetzt werden.

Pitschmann (2001, S. 13 ff.) führt als Auswahlkriterien für externe Materialien, die bei einem solchen Kriterienkatalog zu berücksichtigen sind, unter anderem folgende Punkte an:

- Kontext
  - Herkunft des Dokumentes
  - Beziehung zu anderen Dokumenten
- Inhalt
  - Objektivität
  - Korrektheit
  - Autorität des Urhebers
  - Einzigartigkeit/Originalität
  - Vollständigkeit
  - Umfang
  - Aktualität
  - Zielgruppe
- Zugang
  - Organisation der Ursprungs-Web-Site
  - Navigationsaspekte
  - berücksichtigte Standards
  - Unterstützung des Benutzers
  - Nutzungsbedingungen
  - Rechtlicher Rahmen

Darüber hinaus muss jedoch auf den Bedarf der Nutzer geachtet werden. Ein Dokument, das nicht den gegebenen Kriterien entspricht, aber dennoch äußerst nützlich ist, sollte dennoch in den Bestand übernommen werden, wie dies bei Anschaffungswünschen auch praktiziert wird.

Kriterium 7 *Der Bestandsaufbau muss nach offenen und klar ersichtlichen Auswahl- und Qualitätskriterien erfolgen.*

2.3.3 Digitalisierung

Im Falle der Digitalisierung wird konventionell vorliegendes Material in eine digitale Form überführt und so zur Verfügung gestellt. Dies kann auf verschiedene Weisen erfolgen, bedingt aber manuellen Einsatz und somit einen hohen Aufwand, der sich auch rechnen muss. Darum ist besonders hier auf die ökonomischen Gesichtspunkte zu achten.

Beispiel für eine reine Digitalisierung, die konventionelles (Buch-)Material 1 : 1 wiedergibt, ist die digitalisierte Göttinger Gutenbergbibel[9]. Hierbei kam es auf die original- und detailgetreue Wiedergabe des Dokumentes an und nicht auf den Mehrwert, den ein maschinenlesbarer

[9] `http://www.gutenberg-digital.de/`

Volltext bietet, sowie auf die Sicherung des Originals und dessen Zugänglichmachung für eine breite, uneingeschränkte Öffentlichkeit. Das Projekt kann durch seine Alleinstellung auch nicht im Rahmen der Aktivitäten einer digitalen Bibliothek gesehen werden.

Unter dem Namen *Projekt Gutenberg*[10] fungiert ein Projekt, das gemeinfreie Texte im digitalen Volltext zur Verfügung stellt. Die „Digitalisierung“ erfolgt jedoch durch manuelles Abtippen gedruckter Quellen. Hier stellen die nicht kontrollierbare Sicherheit dieser Quellen (Editionsqualität) und die naturgemäß gegebene Unzuverlässigkeit des Abtippens (Tippfehler) erhebliche Fehlerquellen und Unsicherheitsfaktoren dar. Da es sich hier um ein freiwilliges Projekt handelt, treten diese Faktoren deutlicher zu Tage. Durch geeignete Maßnahmen können diese Negativfaktoren aber geregelt werden.

Kriterium 8 *Digitalisierung von konventionellem Material kann je nach Bedarf eine unterstützende oder vorangige Methode des Bestandsaufbaus sein.*

2.4 Sicherheit

Was einerseits die große Chance an digitalem Material ist, wird unter einer anderen Sichtweise zu einem Problem: seine Körperlosigkeit und die damit einhergehende nahezu unbegrenzte Möglichkeit es neu zu formen, abzuändern oder, negativ ausgedrückt, zu manipulieren.

Die in einem gedruckten Werk hinterlegte Information kann de facto nicht mehr abgeändert werden: es existieren zu viele authentische Kopien davon. Zwar ist die Zahl der Kopien bei digitalen Materialien in der Regel weitaus größer, doch kann hier nur schwer bewiesen werden, welches Exemplar das Original ist und welches eine geänderte Version.

Zu verlockend ist die Möglichkeit, auftretende Tipp-, Schreib- oder sonstige Fehler bei Netzpublikationen zu korrigieren. Solch eine Korrektur bringt jedoch das Problem mit sich, dass sich das Material, das anderweitig vielleicht schon als Quelle dient, unbemerkt ändern kann.

Möglichkeiten um die Authentizität eines Dokumentes zu gewährleisten sind die Techniken der digitalen Signatur und der Prüfsumme.

Bei Prüfsummen handelt es sich um sogenanntes *„hashing“*. Mit nicht-umkehrbaren mathematischen Funktionen wird eine Summe ermittelt, die als „Fingerabdruck“ der geprüften Datei gilt und kein zweites Mal auftreten kann. Ändert sich die Datei, so ist auch die Prüfsumme eine andere. Der Standard für Prüfsummen ist derzeit der *„message digest algorithm, version 5“*, kurz MD5 genannt, der von RSA Data Security entwickelt und in RFC 1321 (Rivest und RSA Data Security 1992) definiert wurde.

Das Beispiel in Tafel 2.1 illustriert verschiedene MD5-Summen für jeweils drei recht ähnliche Inhalte einer ASCII-Datei.

Ein weitergehender Schritt zur Sicherstellung von Integrität, aber auch Authentizität, ist die *digitale Signatur*. Sie stellt nicht nur durch eine Prüfsumme sicher, dass ein Dokument nachträglich nicht verändert worden ist, sie kann auch garantieren, dass diese Signatur von einer bestimmten Person stammt.

Grundlage für dieses Verfahren ist das Private-/Public-Key-Verfahren, welches durch Programme wie *GNU Privacy Guard* (GPG) oder *Pretty Good Privacy* (PGP) implementiert[11] wird und im OpenPGP-Standard nach RFC 2440 (Callas u. a. 1998) definiert ist.

[10] `http://www.gutenberg.org/`; deutsches Projekt unter `http://wwww.gutenberg2000.de/`

[11] `http://www.gnupg.org/` und `http://www.pgp.com/`

Datei-Inhalt	MD5-Summe
Dies ist ein Test	`02feebca49271f25f0e5ff20ced48a8b`
Dies ist kein Test	`be32cdeda56a390587fe4721786c86ca`
Dies ist kein test	`95913621fdb2889fcbd3222b5c49cc4a`
Bibliohtek	`04a3104b209517533ca98ec6be18a39`
Bibliothek	`7ba584bf767793c2c5d793a436fc4fa3`
bibliothek	`27d6bf2ce08fb445130f083c362fda1e`

Tafel 2.1: Prüfsummenbeispiel mit MD5-Hashwerten

Durch den Einsatz eines privaten Schlüssels bei der Prüfsummenbildung wird nicht nur die Einmaligkeit der Prüfsumme gewährleistet, sondern auch der Urheber der Prüfsumme eindeutig gekennzeichnet. Diese digitale Unterschrift trägt somit den Charakter einer handschriftlichen Signatur.

Eine knappe, gut verständliche und ausgewogene Darstellung dieses Themas findet sich bei Ashley u. a. (2000).

Kriterium 9 *Die Integrität und Originalität der Dokumente muss durch zuverlässige Verfahren wie digitale Signaturen oder Prüfsummen sichergestellt werden.*

2.5 Erschließung und Retrieval

Unter *Erschließung* sind jene Tätigkeiten zu verstehen, die dem Benutzer einen Zugang zu den gespeicherten Dokumenten ermöglichen. Erschließung im konventionellen Bereich umfasst das Anlegen einer Titelaufnahme, also Erfassung der bibliografischen Daten, sowie die inhaltliche Beschreibung durch klassifikatorische und/oder verbale Erschließung mittels Systematiken oder Thesauri.

Dieses aufwändige Verfahren wird in konventionellen Bibliotheken intellektuell bewerkstelligt, da das Material keinen anderen Weg zulässt.

Retrieval umfasst die Recherche in erschlossenen Beständen und das Wiederauffinden von Dokumenten, respektive Information. Zu den Pflichten einer digitalen Bibliothek gehört auch die Bereitstellung geeigneter Schnittstellen und Werkzeuge, um eine professionelle Recherche in ihrem Datenbestand zu ermöglichen.

2.5.1 Volltextindexierung

Die Erschließung von Dokumenten durch Volltextindexierung ist eine äußerst verlockende und vielversprechende Möglichkeit. Wird zum Beispiel beim *Regelwerk für den Schlagwortkatalog* (DBI 1982 u. ö.) versucht, den Inhalt eines Dokumentes möglichst genau durch wenige aber präzise Schlagwörter wiederzugeben, so bedingt dies zunächst eine intellektuelle Beschäftigung des Verschlagwortenden mit dem Inhalt.

Liegt ein Text aber bereits in digitaler Form vor, so ist es möglich, jedes darin auftretende Wort suchbar zu machen und dies allein durch maschinelle Methoden und mit einer hohen Verarbeitungsgeschwindigkeit. Der Engpass einer intellektuellen Erschließung und der sich daraus ergebende Zeitaufwand würden somit umgangen.

Kriterium 10 *Bei der Bestandserschließung nutzt eine digitale Bibliothek die Möglichkeiten der Volltextindexierung, jedoch nur unterstützend und nicht als einziges Mittel.*

Die Probleme, die sich dabei durch das Auftreten von Synonymen, Homonymen, Polyphemen, etc. ergeben, werden durch eine einfache Volltextindexierung nicht gelöst und bedingen weitere und aufwändige Methoden.

Eine relativ einfache Möglichkeit ist jene der *Keyphrase-Extraction*, bei der versucht wird, häufig auftauchende Wortkombinationen, sogenannte Kern- oder Schlüsselsätze, im Text zu identifizieren und durch diese seinen Inhalt näher beschreiben zu können, wie es das System *KEA* (für k*eyphrase* e*xtraction* a*lgorithm*) versucht (Witten u. a. 1999a; Frank u. a. 1999). Ein weiteres Beispiel für den Einsatz von Kern- oder Schlüsselsätzen ist das System PHIND, das in Abschnitt 3.6.4 auf Seite 51 demonstriert wird.

Schwierigkeiten bis hin zur Unmöglichkeit ergeben sich bei der Indexierung von nicht textuellem Material. Hierzu zählt auch schon die grafische Darstellung einer Seite, die z. B. eingescannt wurde. Das Retrieval von Bildern oder gar Video- oder Tonsequenzen bringt noch viel weitergehende Hürden mit sich.[12] Ein Beispiel zur Möglicheit des Bildretrievals findet sich in Abschnitt B auf Seite 64.

2.5.2 Metadaten

Metadaten, wörtlich „Daten über Daten“, sind all jene Informationen, die ein Dokument inhaltlich wie auch formal beschreiben. In klassischer Weise sind dies die Katalogeinträge konventioneller Bibliotheken.

Einfach angewandt klären Metadaten formale Fragen wie „Wer ist der Autor?“, „Wie lautet der Titel?“, „Von wann ist dieses Dokument?“. Metadaten können aber auch Angaben zum Inhalt wie Schlagwörter, Systemstellen oder Kurzreferate umfassen.

Als De-Facto-Standard zur Erfassung von Metadaten hat sich inzwischen das *Dublin Core Metadata Element Set* (Dublin Core 1999) entwickelt, welches 15 Felder festlegt, die zur Beschreibung einer Ressource herangezogen werden können, von denen jedoch keines zwingend vorgeschrieben ist:

Title Der Titel des Dokuments

Creator Der hauptsächliche Urheber des Dokuments

Subject Das Thema des Dokuments

Description Eine kurze Beschreibung des Inhaltes: Kurzreferat

Publisher Der Verleger oder Anbieter des Dokuments

Contributor Weitere Urheber oder beitragende Personen

Date Datum

Type Typ des Dokuments: Text, Bild, Ton, etc.

[12] Das ABC-System zum Notensatz ⟨`http://www.gre.ac.uk/~c.walshaw/abc/`⟩ etwa ist textbasiert und würde auf diese Weise das Retrieval vereinfachen. Zur Praxis des „tune retrieval“ siehe McNab u. a. (1996)

Format Speicherformat als MIME-Typ angegeben[13]

Identifier Eindeutige Identifikation des Dokuments (URL, ISBN, DOI)

Source „Quelle“, z. B. Original, von dem das aktuelle Dokument abgeleitet ist

Relation Beziehung zur Quelle: Übersetzung, Bearbeitung, etc.

Language Sprache, in der das Dokument abgefasst ist

Coverage Der Gültigkeitsbereich des Dokumentes bzw. seines Inhaltes. Z. B. „1939 – 1945“ oder „Drittes Reich“

Rights Urheberrechtsvermerk

Das *Dublin Core Metadata Elemet Set* definiert dabei allerdings nur die Felder, durch die eine Ressource beschrieben werden kann. Zwar existiert mit dem *DC Type Vocabulary* ein Ansatz zur Definition, wie diese einzelnen Felder anzugeben sind, jedoch ist dieser zum einen stark anglo-amerikanisch orientiert und lässt zum anderen immer noch viele Fragen offen.

So sind all jene Fragen, die die einschlägigen bibliothekarischen Regelwerke klären, unbeantwortet. Etwa in welcher Weise der Name eines Autors, das Datum oder der Titel anzugeben sind.

Diesem Problem begegnet das *Qualified Dublin Core* (Dublin Core 2000), das die Angabe der Metadaten genauer spezifiziert.

Bei der Erfassung von Metadaten stoßen automatische Methoden aufgrund der Unstrukturiertheit der zu erfassenden Daten hart an ihre Grenzen (vgl. Yeates 1999). Dies muss daher mit gutem Recht intellektueller Arbeit vorbehalten bleiben.

Kriterium 11 *Der Bestand einer digitalen Bibliothek wird mit bibliografischer Sorgfalt erschlossen und Titeldaten werden zuverlässig und normgerecht erfasst. Eine inhaltliche Erschließung muss ebenfalls nach den Regeln der Kunst erfolgen.*

2.5.3 Indizes

Eine Suche in Datenbeständen wird in der Regel nie über den eigentlichen Bestand, sondern über daraus gebildete Indizes erfolgen. Bei der Erstellung der Indizes kann auch auf sinnvolle Maßnahmen Wert gelegt werden, wie die Eliminierung von Stoppwörtern[14] oder, vor allem bei aus Volltexten gebildeten Indizes, eine Reduktion der Wörter auf ihre Grundform (z. B. bei Flexionsformen).

Die Repräsentation des Datenbestandes und der Zugang dazu bestehen alleine durch Indizes. Daten, die darin nicht auftauchen, sind auch nicht suchbar. Da sich Indexierung erheblich auf die Leistung beim Erstellen der Indizes, bei deren Abfrage und auf deren Größe auswirkt, werden teils nicht alle Daten indexiert und suchbar gemacht. So stellen vor allem Volltextindizes hohe Anforderungen an die Leistungsfähigkeit der verarbeitenden Rechner.[15]

[13] Die *Multipurpose Internet Mail Extensions* waren dazu gedacht, durch Angabe der Inhaltstypen textuelle und nicht-textuelle Inhalte in elektronischer Post zu ermöglichen und sind inzwischen der De-Facto-Standard zur generellen Angabe von Inhaltstypen. Vergleiche hierzu Borenstein und Freed (1993) sowie Freed und Borenstein (1996).

[14] Häufig auftretende Wörter, die jedoch keinen Sinngehalt besitzen, wie Artikel, Präpositionen, etc.

[15] Ein Beispiel für Größe und Geschwindigkeit findet sich in Abschnitt 3.6.3 auf Seite 49.

Operator	Bedeutung	Funktion
`AND`	logisches Und	Beide Bedingungen müssen gegeben sein.
`OR`	logisches Oder	Eine von beiden Bedingungen muss gegeben sein.
`NOT`	logisches Nicht	Die zweite Bedingung darf nicht gegeben sein.
`XOR`	exklusives Oder	*Entweder* die erste *oder* die zweite Bedingung muss gegeben sein.

Tafel 2.2: Die boole'schen Operatoren

Grundsätzlich ist zu erwarten, dass ein Index zu den Metadaten besteht, der wenigstens die wichtigsten wie Autor, Titel, etc. umfasst. Der genaue Aufbau hängt in jedem Fall vom aktuellen Bedarf und der Suchweise der Nutzer ab, ob beispielsweise schwerpunktmäßig über die Autoren, den Titel, die Schlagwörter oder den Volltext gesucht wird.

Indizes aus Volltexten sollten aus oben geschilderten inhaltlichen Gründen von den aus Metadaten gebildeten getrennt sein, so dass diese auch getrennt abgefragt werden können. Eine Einbeziehung einiger oder aller Indizes in eine Suche sollte möglich sein.

Kriterium 12 *Aus den Daten und Metadaten sind je nach Bedarf und Sinn unterschiedliche Indizes zu bilden, die sowohl getrennt als auch gemeinsam abfragbar sind. Indexiert werden sollen alle für den jeweiligen Bedarf relevanten Daten und Metadaten.*

2.5.4 Retrieval

Boole'sche Suche

Die *boole'sche Suche* basiert auf der von dem englischen Mathematiker George Boole entworfenen Logik (*Boole'sche Algebra*), gemäß derer sich Zustände durch die Aussagen *wahr* oder *falsch* beschreiben lassen. Diese Aussagen lassen sich durch die drei Operatoren `AND`, `OR` und `NOT` definieren. Weniger häufig vertreten ist ein erweiterter Operator `XOR`, der ein ausschließliches Oder („entweder – oder") symbolisiert.

Durch das boole'sche Prinzip kann ein Objekt durch die In-Beziehung-Setzung seiner Eigenschaften durch die 4 Operatoren beschrieben werden, wie in Abbildung 2.1 dargestellt. Dabei werden jeweils zwei Bedingungen (Eigenschaften) in der Reihenfolge von links nach rechts gegeneinander verifiziert. In einer Suche grenzt man damit eine Menge von Elementen gegen die Gesamtmenge ab. Diese Suchmethode lässt sich auch als *logische Suche* bezeichnen.

Die boole'sche Suche stellt im professionellen Information Retrieval den de-facto-Standard dar und ist somit in allen Retrievalsystemen vorauszusetzen.

„Ranked Query"

Da das boole'sche Suchprinzip durch die Einbeziehung oder den Ausschluss bestimmter möglicher vorhandener Merkmale eines oder mehrerer Objekte zu einer Ergebnismenge führt, setzt diese Methode die Kenntnis wie auch die Existenz dieser Eigenschaften voraus.

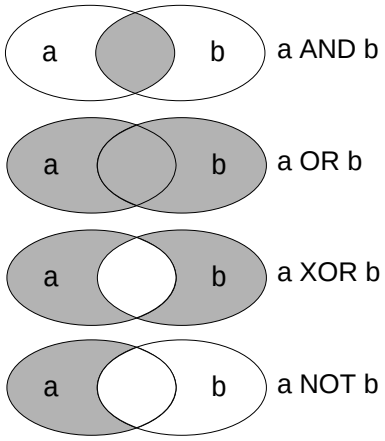

Abbildung 2.1: Prinzip der boole'schen Logik

Titel	rank-Value
Drei Chinesen *mit dem* Kontrabass	100 %
Vier Chinesen *mit dem* Kontrabass	66 %
Drei Schweizer *mit dem* Kontrabass	66 %
Zwei Berliner *mit dem* Kontrabass	33 %
Asiatisches Streicher-Trio	0 %

Tafel 2.3: Beispiel eines gestaffelten Ergebnisses

Die Suche nach `drei AND chinesen AND kontrabass` ist vergeblich, wenn es kein Objekt (im Falle einer Bibliothek kein Dokument) gibt, welches diese drei Eigenschaften enthält (z. B. der Titel).

Die *ranked query* dagegen geht davon aus, dass nicht alle Eigenschaften notwendiger Weise vorhanden sein müssen und liefert daher als Ergebnismenge eine Staffelung („ranking“) einzelner Objekte zurück, die den vorgegebenen Kriterien *bestmöglichst* aber nicht unbedingt vollständig entsprechen. Ordnendes Kriterium ist das Maß, in welchem die Elemente der Ergebnismenge mit einer gegebenen Liste von Eigenschaften übereinstimmen. Ein Beispiel hierfür liefert Tafel 2.3 für die Suche `drei chinesen kontrabass`. Die Stoppwörter erscheinen kursiv.

Ranked queries werden besonders bei Web-Suchdiensten eingesetzt und haben von daher zweifelhafte Bekanntheit erlangt.[16]

Grundsätzlich ist es möglich, gestaffelte Suchen mit boole'schen zu koppeln, z. B. `(drei chinesen kontrabass) NOT konzert`, was aber eine recht aufwändige Verarbeitung der Abfrage bedingt.

Alternative Suchmethoden

Als **geführte Suche** kann das Browsen durch aus den Dokumenten extrahierte Kernsätze bzw. KWIC-Indizes sein. Für dieses Vorgehen bietet sich der Begriff des *„guided keyword“* an, wie

[16] Die Verfälschung der Suchergebnisse durch Sponsoring führte oftmals zu Kritik und Zweifeln. Siehe Becker (2001).

es sich in der PHIND-Suche (siehe Abschnitt 3.6.4 auf Seite 51) darstellt. In weitergehender Weise wird dieses Prinzip durch das System *Phrasier* (Jones 1998; Jones und Staveley 1999) implementiert.

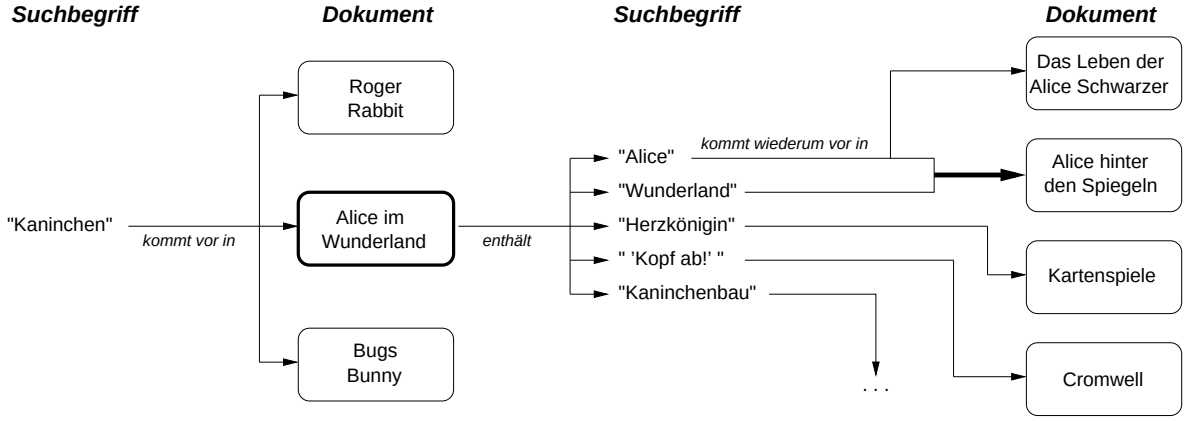

Abbildung 2.2: Prinzip der geführten Suche („*guided keyword*“)

Die geführte Suche nimmt ein Suchkriterium, in diesem Fall einen Suchbegriff, als Ausgangspunkt, auf Grund dessen eine Ergebnismenge zurückgeliefert wird. Im in Abbildung 2.2 dargestellten Beispiel geschieht dies durch ein einzelnes Wort. Bei der Verwendung von mehreren Suchbegriffen kann die Bildung der Ergebnismenge durch verschiedene Methoden erfolgen, wie z. B. die oben beschriebenen boole'schen und gestaffelten Suchmethoden. Der besondere Wert der geführten Suche besteht nun darin, dass weitere mögliche Suchbegriffe, die in anderen Dokumenten vorkommen, dargestellt werden und anhand derer die Suche fortgesetzt werden kann. So könnte die ursprüngliche Suche nach einem Kaninchen auch zu einer Biografie von Alice Schwarzer oder einem Schauspiel über Oliver Cromwell führen.[17]

Einen weiteren und sinnvollen Einstieg bietet die klassische **Registersuche**, auf die auch bei allen Volltext- und Metadatenindizes nicht verzichtet werden kann.

Hinter einer Registersuche steht das klassische Blättern bzw. „browsing“ in einer alphabetisch oder systematisch geordneten Liste von Begriffen, analog zum Aufbau alter Kartenkataloge oder dem Zugang zu gedruckten Wörterbüchern und Lexika, etc. Obwohl dies recht antiquiert anmutet, verfügen die großen Hosts immer noch über diese Möglichkeit.[18]

Die Registersuche ermöglicht auf diese Weise einen Überblick über die auftretenden Begriffe, bevor eine Ergebnismenge zurückgeliefert wird und erweist sich dann als hilfreich, wenn die Existenz eines Suchbegriffes oder dessen genaue Schreibweise nicht sicher ist, bzw. wenn in der Aufnahme ein Schreibfehler enthalten ist (Abbildung 2.3 zeigt ein Beispiel aus dem Stuttgarter Regionalkatalog[19], samt Tippfehler im vorletzten sichbaren Eintrag). Überhaupt erlauben Register einen ungezwungenen Überblick über das Vorhandene.

Eine hinsichtlich der Benutzerfreundlichkeit sinnvolle Ergänzung zu allen Suchmethoden ist das als *„fuzzy logic“* bezeichnete Konzept, offensichtliche und vermutliche Tippfehler bei der Suchanfrage automatisch zu korrigieren oder darauf hinzuweisen.

[17] Ein sinnvolleres und praktisches Beispiel findet sich in der Demonstration in Abschnitt 3.6.4 auf Seite 51.

[18] *Dialog* ⟨`http://www.dialog.com/`⟩ und *STN* ⟨`http://www.stn-international.de/`⟩etwa bieten dies über die Funktion `expand` an.

[19] `http://www.biss.belwue.de/cgi-bin/bissform.cgi`

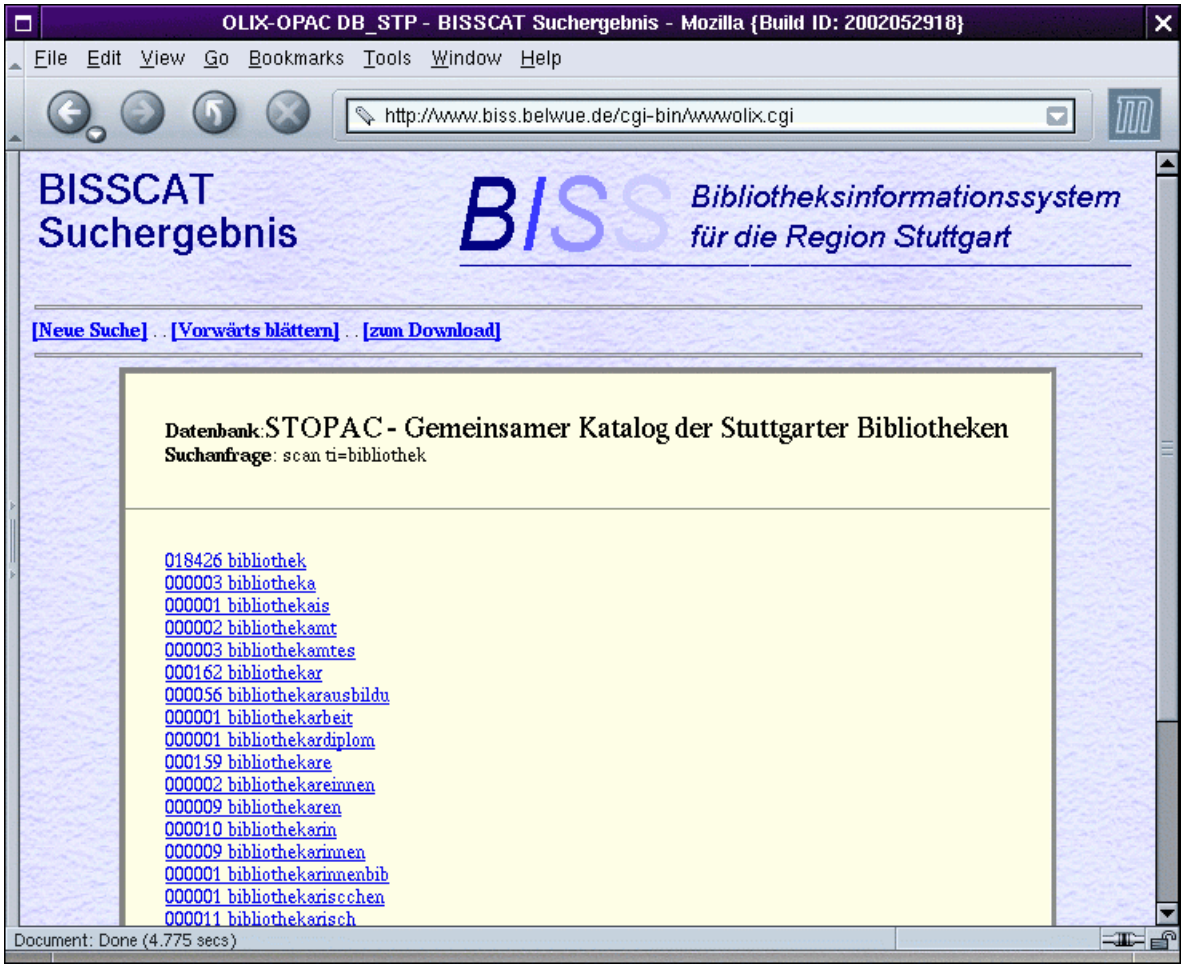

Abbildung 2.3: Registersuche im BISSCAT

Kriterium 13 *Die Bestände einer digitalen Bibliothek sollen durch boole'sche Suchmethoden abfragbar sein. Weitere, sich ergänzende Suchmethoden sind unbedingt wünschenswert.*

2.5.5 Retrieval-Oberfläche

Grundlegend existieren drei Arten von Retrieval-Oberflächen, die sich hinsichtlich ihrer Nutzung und Funktion unterscheiden. Dies sind kommandoorientierte, formulargesteuerte und grafische Oberflächen oder *Interfaces*.

Kommandoorientierte Oberflächen sind die älteste Art des Zugangs und basieren auf der Terminal-Technik, bei der dem Benutzer eine lokale oder entfernte („*remote*“) Kommandozeile zur Verfügung gestellt wird, über die er seine Suchkommandos eingeben kann und die Ergebnisse angezeigt bekommt.

Formulargesteuerte Oberflächen kamen mit der Entwicklung des World Wide Web auf und stellen zur Formulierung der Abfrage Eingabemasken zur Verfügung. Dieses Prinzip ist durch seine Selbsterklärlichkeit deutlich benutzerfreundlicher, jedoch immer wieder weniger leistungsstark wie eine kommandoorientierte Oberfläche.

Grafische Retrieval-Oberflächen sind äußerst selten und in wenigen Fällen sinnvoll. So werden sie vor allem bei der Recherche grafischer Objekte wie Bilder eingesetzt;[20] in chemischen Datenbanken ist die Suche nach Molekülstrukturen über solche Systeme realisiert.

Bei der Recherche textueller Daten kommt dieses Prinzip beim sogenannten *„vector space model“* zum Tragen. Dabei werden die Beziehungen der Suchbegriffe grafisch modelliert und anschließend als Suche formuliert. Anschaulich ist dies bei Jones u. a. (1999) dargestellt.

2.6 Literatur- und Informationsversorgung

Nach der Definition aus Abschnitt 1.2.6 auf Seite 6 bleibt nun die Frage nach dem Unterschied zwischen Literatur- und Informationsversorgung.

Ein Punkt an Mehrwert ist über die Bereitstellung des Materials hinaus dessen Erschließung mit oben genannten Möglichkeiten. Bei einem systematisch und inhaltlich orientierten Bestandsaufbau, der vorrangig externes Material heranzieht und so seinem Bestand ein Auswahlverfahren voranstellt, ergibt sich aus den gespeicherten Dokumenten eine inhaltliche Beziehung alleine daraus, dass diese gemeinsam vorhanden sind.

Einen bedeutenden Faktor nimmt dabei Information ein, von deren Existenz ein Nutzer keine Kenntnis hat. Ein Kunststück bibliothekarischer Arbeit bleibt es also nach wie vor den Nutzer mit Information zu versorgen, die er braucht, aber nicht vermutet hat.

Kriterium 14 *Relevante Information muss für den Benutzer verfügbar sein, bevor er diese benötigt.*

Dieser Umstand *proaktiven Bestandsaufbaus* bedingt natürlich eine enge Zusammenarbeit mit den Nutzern sowie eine genaue Kenntnis deren Informationsbedarfs.

Von überaus großer Bedeutung ist und bleibt aber auch Information, die nicht lokal vorhanden ist oder für die der Nutzer Hilfestellung braucht. In konventionellen Bibliotheken übernimmt diese Aufgabe die Information bzw. Auskunft. Dieses wertvolle Instrument darf auch in einer digitalen Bibliothek nicht fehlen, nicht zuletzt um in der sonst ausschließlichen Mensch-Maschine-Kommunikation auch den persönlichen Kontakt zu ermöglichen.

Kriterium 15 *Kontaktmöglichkeiten mit der Funktion eines Helpdesk oder einer Auskunft müssen gegeben sein.*

Unterschiede im Typus der digitalen Bibliothek ergeben sich aus der Art des Bestandes und des auf ihm aufsetzenden Mehrwertes. Hier lassen sich in grober Gliederung drei Arten von Servern definieren, die das Kernstück einer digitalen Bibliothek bilden.[21]

Auf unterster Ebene stehen sogenannte **Dokumentserver**, deren alleinige Aufgabe die Speicherung und Bereitstellung von Dokumenten ist.

Die nächste Stufe der **Informationsserver** bezieht in den Dokumentbestand zum einen weitergehende Informationen wie die genannten Volltextindizes und Metadaten mit vor allem inhaltlichen Angaben mit ein und stellt durch die Vorauswahl der Dokumente und deren systematische (und „synoptische“) Sammlung eine höhere Qualität mit einem höheren Basisgehalt

[20] Zur Darstellung des Bildretrieval siehe Abschnitt B auf Seite 64.

[21] *„Server“* meint hier die Komponente, welche den Bestand enthält und um die herum sich eine digitale Bibliothek aufbaut.

an Information dar. Hierzu zählen auch Informationen zu den Beziehungen der Dokumente untereinander (z. B. welches Dokument in welchem zitiert wird).

Auf der höchsten Stufe steht der **Wissensserver**. Dieser ist in der Regel theoretischer Natur und würde alle Informationen enthalten, die ein Nutzer während seiner Arbeit hinzugezogen hat, das Ergebnis dieser Arbeit wie auch Informationen darüber, welches der benutzten Dokumente in welcher Weise nützlich oder unnütz war.

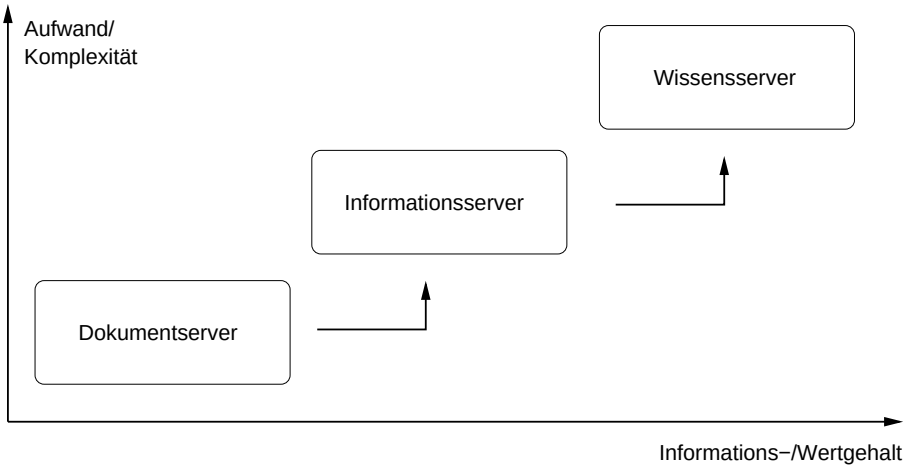

Abbildung 2.4: Verschiedene Servertypen

Abbildung 2.4 illustriert die Staffelung dieser verschiedenen, größtenteils theoretischen, Typen. Die Verwirklichung dieser Ansätze scheitert weniger an der technischen Machbarkeit als an dem hohen Aufwand, der nicht nur dem Betriebspersonal, sondern vor allem den Nutzern abverlangt wird.

Kriterium 16 *Eine digitale Bibliothek verbindet bibliothekarische und dokumentarische Arbeit und leistet so einen weitgehenden Schritt zum Mehrwert ihres Bestandes.*

2.7 Zugang und Distribution

Digitale Bibliotheken handeln mit Objekten, die an keinen Ort und an keine feste Anzahl von Exemplaren gebunden sind. Insofern muss dieses Potential wo nur irgend möglich ausgeschöpft werden. Die weltweite Infrastruktur des Internet bietet dazu die besten Voraussetzungen. Das Angebot einer digitalen Bibliothek ist somit weder an Ort noch an Zeit gebunden.

Kriterium 17 *Eine digitale Bibliothek stellt ihre Dienste ortsungebunden und zeitlich uneingeschränkt zur Verfügung.*

Wenn Zugang und Inhalte wie gefordert frei sein sollen, so muss auch der Zugang zu diesem Dienst ohne Hürden möglich sein. Am besten garantiert dies der Einsatz offener und freier Protokolle, also lizenzfreier Standards. Auf der Seite des Anbieters können somit schon vorhandene Komponenten und Erfahrungen in den Dienst einfließen, auf Nutzerseite sind keine proprietären Schnittstellen notwendig. Der De-Facto-Standard für netbasierte Dienste ist

das *World Wide Web*: das HTTP-Protokoll bietet weitgehende Möglichkeiten zum Datenaustausch.[22]

Kriterium 18 *Der Zugang zu einer digitalen Bibliothek erfolgt über offene Protokolle und mit Hilfe offener Standards. Derzeit ist dies die Technik des* World Wide Web.

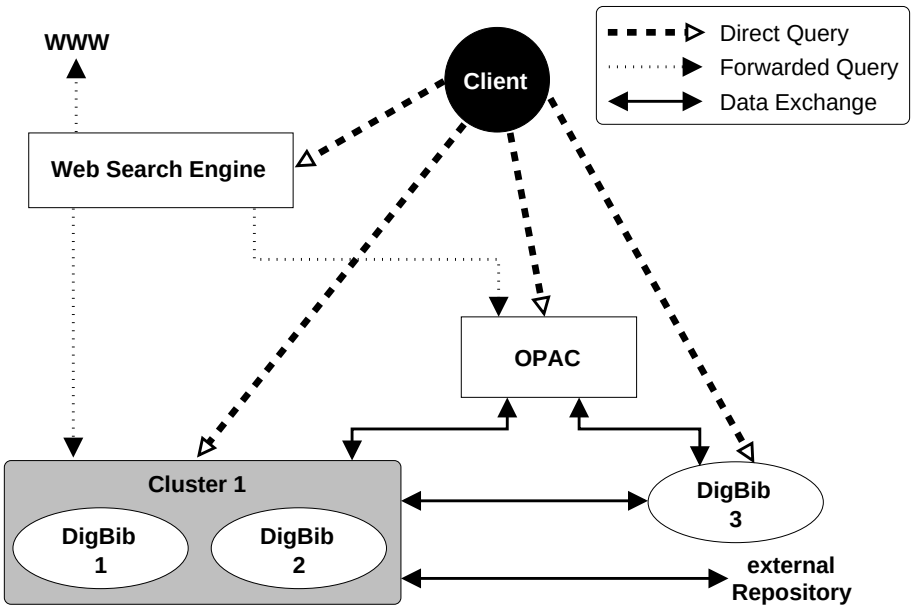

Abbildung 2.5: Interoperabilitätsbeispiel

Die darüber hinaus existierenden Möglichkeiten, automatisiert Daten mit anderen Anwendungen auszutauschen, müssen in vielen Fällen als gegeben vorausgesetzt werden. Allen voran ist dies das *Open Archives Initiative Protocol for Metadata Harvesting*[23], das als Meta-Protokoll HTTP-basiert ist und ein XML-kodiertes Schema zum Austausch von Metadaten zwischen Publikationsserver und Clients definiert.

Der ältere, inzwischen in der dritte Fassung vorliegende, Standard *Z39.50* (ANSI 1995) definiert ein komplettes Protokoll zur Formulierung von Suchanfragen (*„queries“*) und deren Antworten. Ziel war es, fremde Retrieval-Systeme aus eigenen heraus abfragen zu können, ohne dass deren Funktionsweise dem Endanwender bekannt sein muss. Die Übersetzung der Anfragen und die Kommunikation mit dem Zielsystem erfolgte über Z39.50.

Eine weitere Möglichkeit bietet der Einsatz der *„Common Object Request Broker Architecture“* (CORBA), die auf Applikationsebene arbeitet und Funktionsaufrufe einer Anwendung auf mehrere Komponenten aufteilt, die auch über Netzwerke verteilt sein können. Ein Beispiel zum Einsatz von CORBA liefert Bainbridge u. a. (2001).

Vielversprechendes Entwicklungspotential bieten hier auch die zur Zeit viel diskutierten *Web Services*.[24] Abbildung 2.5 illustriert das Prinzip der Interoperabilität.

[22] Die Pflege und Weiterentwicklung liegt in der Hand des W3C ⟨`http://www.w3.org/Protocols/`⟩.
[23] Siehe dazu die Web-Site der OAI ⟨`http://www.openarchives.org/`⟩.
[24] Als Projekt am W3C angesiedelt ⟨`http://www.w3.org/2002/ws/`⟩.

Kriterium 19 *Schnittstellen zum automatisierten Datenaustausch mit anderen Anwendungen und Diensten müssen auf Basis offener und freier Protokolle vorhanden sein.*

Problematisch bei netzwerkbasierten Diensten ist immer die Verfügbarkeit des Netzwerkes und der Kostenfaktor. Dieser war einer der ausschlaggebenden Punkte, die *New Zealand Digital Library* einzurichten (Witten u. a. 1999b, S. 482). Verfügbarkeit und Transfergeschwindigkeit des Netzwerkes stellen in Nicht-Industrieländern nach wie vor einen Engpass beim Transfer größerer Datenmengen dar, wie auch Benutzer mit immer noch weit verbreitetem Modem-Anschluss ebenfalls sehr schnell die Grenze des Zumutbaren erreichen (vgl. Zivadinovic 2001).

Die Möglichkeit, unabhängig von einem Netzwerkzugang auf die Bestände einer digitalen Bibliothek zugreifen zu können, sollte daher gegeben sein. Die Distribution der kompletten oder teilweisen Bestände auf Datenträgern (etwa CD-ROM oder DVD) sollte daher in Betracht gezogen werden.

Kriterium 20 *Eine digitale Bibliothek sollte wie irgend möglich auch offline verfügbar sein.*

2.8 Trägerschaft

Wie der bisherigen Darstellung und auch den vordefinierten Schwerpunkten hervorgeht, decken sich die Tätigkeitsfelder von konventionellen und digitalen Bibliotheken nahezu. So ist nach Friedling (2001) heute

> „[. . .] der physische Besitz von Information nicht mehr Hauptziel des Bibliotheksmanagements. Statt dessen übernimmt die Bibliothek Aufgaben der Informationsvermittlung, indem sie weltweite Zugänge und Zugriffe auf Informationen ermöglicht. Die Bibliothek übernimmt wissensorganisierende und inhaltlich orientierende Funktionen. Sie schafft Mehrwerte durch eigene Informationsprodukte (Datenbanken, Aufbau und Betreuung von Volltextservern, Informationsportalen und Wissensmanagementsystemen).“

Und auch der Wissenschaftsrat (2001, S. 28) äußert sich hierzu:

> „Traditionell ist die Hochschulbibliothek innerhalb der Hochschule sowie im lokalen und regionalen Bereich die zentrale Dienstleistungseinrichtung für die umfassende Nutzung vieler Medienformen in Forschung, Lehre und Weiterbildung.“

Dementsprechend stellt sich das hier dargelegte Konzept als konsequente Weiterentwicklung bibliothekarischer Dienstleistung dar.

Nicht näher wird in dieser Arbeit auf die Personalanforderungen eingegangen, doch lassen sich hier, bei der Realisation durch eine Bibliothek oder ähnliche Einrichtung, bestehende Kompetenzen sinnvoll nutzen.

Kriterium 21 *Eine digitale Bibliothek ist nur ein weiteres Angebot einer bestehenden Institution des BID-Bereiches und nutzt so bestehendes Potenzial.*

2.9 Klientel

Folgerichtig aus Abschnitt 2.8 abgeleitet, gilt als Klientel einer digitalen Bibliothek zunächst die der tragenden Einrichtung.

Digitale Bibliotheken sind jedoch, wie schon genannt, nicht an Orte oder Zeiten gebunden. Ihr Angebot besteht rund um die Uhr und auch rund um den Globus. Jedes Individuum mit einem entsprechenden Anschluss an das Internet könnte somit auf die digitale Bibliothek zugreifen, wie auch prinzipiell jeder eine öffentlich zugängliche Bibliothek benutzen darf.

Überhaupt sind die erbrachten Dienstleistungen zu wertvoll, um sie in ihrer Nutzung auf eine kleine, spezielle Zielgruppe einzuschränken. In dem Gedanken, dass Information frei sein muss, soll auch der Zugriff auf die Dienstleistungen einer digitalen Bibliothek allen Menschen offen stehen.

Dass eine weltweite Klientel aber auch das Maß der zu bewältigenden Arbeit sprengt, ist offensichtlich. Auch wenn man prinzipiell die Weltbevölkerung als Kunden hat, so bleibt die *Zielgruppe* doch schon alleine aus sprachlichen Gründen eine wesentlich spezifischere.

Kriterium 22 *Grundsätzlich muss jeder Mensch, dem der Zugang technisch möglich ist, potentieller Kunde einer digitalen Bibliothek sein können. Ihr Profil und ihre Dienstleistungen sind jedoch auf eine bestimmte Klientel ausgerichtet, in der Regel die der Trägereinrichtung.*

2.10 Rechtliche Aspekte

Nicht zuletzt zu berücksichtigen bleiben auch die Rechtsnormen, die für den Betrieb einer digitalen Bibliothek relevant sind. Da das Internet aus den Anbietern netzbasierter Dienste durch seine weltweite Struktur gleichzeitig *„Global Players“* macht, sind hier nicht nur nationale sondern auch internationale Gesetze und Vorschriften zu beachten.

Die Beschäftigung mit den entsprechenden relevanten Regelungen, die bei den verschiedensten Fällen zu berücksichtigen sind, würden schon das Maß einer Arbeit sprengen, die sich alleine mit diesem Thema beschäftigt. Daher soll hier nur ein kurzer Überblick über die wichtigsten Gesetze Platz finden.

Zur genauen Klärung des rechtlichen Standes und der geltenden Vorschriften ist in jedem Fall die Unterstützung eines Experten notwendig.

Generell sind bei der Nutzung von Materialien die **Urheberrechte** zu beachten, die je auf nationaler Ebene geregelt sind und durch das Welturheberrechtsabkommen eine internationale Dimension bekommen haben. Hier werden primär die Rechte des Urhebers an seinem Werk geregelt, aber auch, welche Nutzung geschützter Materialien in welchem Umfang erlaubt ist.

Leider unterliegen konventionelle und digitale Materialien im Rechtsverständnis immer wieder anderen Bedingungen, was Nancy Kranich, Präsidentin der *American Library Association* (ALA)[25], kritisierte (Weeks 2001):

> In Kranich's mind, library-goers should be able to duplicate limited amounts of information for educational purposes. Suppose you want to copy a journal article, quote a section of a book or use a line from a poem, she says. „That is all permitted under the fair-use provision of the copyright law. In the digital arena, fair use has been narrowed to the point of disappearing.“

[25] `http://www.ala.org/`

> „The publishing community does not believe that the public should have the same rights in the electronic world," Kranich says.

In Deutschland ist durch das **Teledienstgesetz** bzw. den **Mediendienstestaatsvertrag** (MDStV) der Status von Dienstbetreibern festgelegt, die vor allem die Regelungen zur Haftbarkeit von Inhalten enthalten.

Weitere Regelungen von Bedeutung sind das **Gesetz gegen die Verbreitung jugendgefährdender Schriften und Medien**, die jeweiligen **Datenschutzgesetze** des Bundes und der Länder, sowie das **Marken- und Namensrecht**.

Eine Sammlung der relevanten Gesetzestexte mit Kommentar findet sich z. B. bei Geppert und Roßnagel (1998).

Teil II

Praktischer Teil

Kapitel 3

Implementierung

Im Folgenden soll die Implementierung einer digitalen Bibliothek mit der *„Greenstone Digital Library Software“* dargestellt werden, wobei die Kriterien aus Kapitel 2 als Maßstab dienen sollen. Dass bereits vorgefertigte Lösungen nicht zu 100 % damit übereinstimmen können, liegt in der Natur der Sache.

Bei einem früheren Projekt an der HBI/HdM Stuttgart fiel die Wahl jedoch deshalb auf die Greenstone-Software, da diese dem damaligen Kriterienkatalog weitgehend gerecht wurde (vgl. Engster u. a. 2001).

3.1 Wahl der Software

Als Software, mit der versucht werden soll, die Konzeption aus Kapitel 2 zu implementieren, wurde die *Greenstone Digital Library Software* (GSDL, im Folgenden als „Greenstone“ bezeichnet) herangezogen. Dieses Paket kommt beim *New Zealand Digital Library Project* sowie bei diversen nicht-kommerziellen Projekten zum Einsatz (vgl. UNESCO 2000).

Greenstone wird an der Fakultät für Informatik der Universität von Waikato[1] in Hamilton, Neuseeland mit Unterstützung der UNESCO[2] entwickelt und kann über das Internet[3] kostenfrei bezogen werden.

Grund der Wahl für die im Folgenden vorgestellte Implementierung waren folgende Punkte:

- die Entwicklung im akademischen Bereich und die Unterstützung durch eine überstaatliche Organisation sichert weitgehend die Weiterentwicklung und Pflege der Software
- es handelt sich um freie Software, die unter den Bedingungen der *GNU General Public License* (GPL)[4] lizenziert ist und kann somit frei benutzt, angepasst und verteilt werden; zudem fallen keine Lizenzgebühren an
- das Paket ist auf UNIX-, Windows- und MacOS X-Systemen lauffähig
- die Software stellt geringe Systemanforderungen

[1] `http://www.cs.waikato.ac.nz`
[2] United Nations Educational, Scientific and Cultural Organization ⟨`http://www.unesco.org`⟩
[3] Von der Web-Site des Greenstone-Projektes ⟨`http://www.greenstone.org`⟩.
[4] `http://www.gnu.org/licenses/gpl.html`

- das Paket ist modular aufgebaut und integriert bereits vorhandene freie Softwarepakete für einzelne Aufgaben und stellt dadurch keine vereinzelte Lösung dar
- die Möglichkeit der Publikation auf CD-ROM ist integriert
- es existieren Schnittstellen zu Techniken wie der Verteilung des Datenbestandes über CORBA oder der Extraktion von Kernsätzen
- das Paket verfügt über eine leistungsfähige Indexierungs- und Retrieval-Engine

Die Verquickung der Punkte „freie Software“ und „modulare, offene Schnittstellen“ erlaubt in Verbindung mit einiger Experimentierfreudigkeit eine weitgehende Anpassung an eigene Bedürfnisse sowie die Einbindung komplett neuer Funktionalitäten.

Greenstone ist daher bei weitergehendem Einsatz als Rahmenwerk zu sehen, das auf den Funktionen einer leistungsfähigen Retrieval-Engine aufsetzt und sich nach dem Baukastenprinzip abspecken oder aufrüsten lässt.

3.2 Systemanforderungen

Die minimalen Anforderungen der Software hängen zwar immer von dem verwendeten Betriebssystem ab, sind jedoch sehr gering gehalten. So läuft die Software bereits schon auf als veraltet geltenden Intel-386-Systemen.

Jedoch ist der Einsatz von Greenstone zweigleisig: zum einen ist sie als serverseitige Anwendung gedacht, zum anderen aber auch als Offline-Version für den heimischen Rechner. Je nach Einsatzgebiet stellt allein schon die Lastverträglichkeit andere Anforderungen.

Im Folgenden steht jedoch der Serverbetrieb im Vordergrund.[5] Als Basis für die Implementierung wurde ein PC-System mit einem Intel-5/686-kompatiblen 300-MHz-Prozessor, 128 MB Arbeitsspeicher und 10 GB Festplattenspeicher verwendet. Als Betriebssystem kam die Debian-GNU/Linux-Distribution mit der Kernel-Version 2.4.17 zum Einsatz.

Als Webserver wurde der Apache in der Version 1.3.24 verwendet. Grundsätzlich ist jeder CGI-fähige[6] Webserver dazu geeignet, ein Greenstone-System zu beheimaten. Die Vorteile des Apache liegen darin, dass er ebenfalls zur Kategorie freie Software gehört, seine Anteile als Marktführer ihn als den besten zur Zeit erhältlichen Webserver ausweisen[7] und er eine große Menge an Anpassungs-, Erweiterungs- und Optimierungsmöglichkeiten bietet, sowie auch auf allen großen Systemplattformen vertreten ist[8].

Desweiteren ist eine funktionstüchtige Perl[9]-Installation notwendig. Perl-Module, die während der Arbeit benötigt werden, aber nicht vorausgesetzt werden können, sind in der Distribution schon enthalten und werden vor den lokalen Modulen geladen. Es stellt somit kein Problem dar, wenn lokal andere Versionen derselben Module vorhanden sind. Für diese Arbeit wurde Perl in der Version 5.6 eingesetzt.

[5] Mit Ausnahme von Abschnitt 3.7 auf Seite 52.

[6] *Common Gateway Interface*: Schnittstelle eines Webservers zu weiteren Programmen, welche die Verarbeitungen von Anfragen aus dem Web übernehmen können. Siehe auch Fußnote 11 auf der nächsten Seite.

[7] nach dem *Netcraft Web Server Survey* ⟨`http://www.netcraft.com/survey/`⟩

[8] Siehe die Homepage des Apache-Webserver-Projektes ⟨`http://httpd.apache.org/`⟩.

[9] Skriptorientierte Programmiersprache, siehe `http://www.perl.com`.

3.3 Installation des Greenstone-Paketes

Der genaue Installationsvorgang ist im *„Installer's Guide“* beschrieben (Witten und Boddie 2002b) und soll hier nicht im Vordergrund stehen. Daher beschreibt dieser Abschnitt vor allem die speziellen Punkte der Installation, die für die hier vorgestellte Modell-Implementierung von Interesse sind.

Ansonsten wird in den nachfolgenden Abschnitten von einer funktionierenden Installation ausgegangen.

3.3.1 Allgemeine Installation

Zur Installation stehen zwei Pakete zur Auswahl: ein binäres, vorkompiliertes Paket, das für die entsprechende Plattform sofort betriebsfertig ist, oder ein Quellcodepaket. Letzteres bietet die Möglichkeit bei der Kompilierung Optionen nach eigenen Wünschen hinzuzufügen oder wegzulassen. Die nennenswertesten Optionen sind CORBA- und Z39.50-Unterstützung sowie die Optimierung über FastCGI (siehe Abschnitt 3.3.2).

Neben den Eigenentwicklungen des Greenstone-Projektes besteht das Softwarepaket aus zahlreichen weiteren Programmen, die alle in der Distribution enthalten sind.

Das Makefile der Quellcodedistribution oder das Installationsskript der Binardistribution bewerkstelligen die notwendigen Aufgaben, die zur Installation aller Komponenten zuständig sind.

Das Basisverzeichnis, in das Greenstone installiert wurde, wird im Folgenden, übereinstimmend mit den Skripten aus der Distribution, mit der Variable `$GSDLHOME` beschrieben.

Unter Unix-Systemen ist vor allem auf die richtig gesetzten Zugriffsrechte für Dateien und Verzeichnisse sowie die auszuführenden Skripten zu achten.

Nach der Installation per Skript oder aus der Quellcodedistribution direkt heraus ist die Konfiguration des Webservers anzupassen. Ein Beispiel hierfür gibt Tafel 3.1.

3.3.2 Optimierung

Greenstone bietet bei einer Installation aus der Quellcode-Distribution heraus die Möglichkeit, die CGI-Anwendung über FastCGI optimiert zu betreiben.

FastCGI[10] ist eine Erweiterung des CGI-Protokolls[11] um die Ausführung serverseitiger Skripten zu beschleunigen. Dazu steht ein API[12] zur Verfügung, das in die zu optimierende Anwendung integriert wird und mit einem entsprechenden Webserver-Modul kommuniziert. FastCGI ist für die verbreitetsten Webserver sowie für eine Vielzahl von Programmiersprachen verfügbar.

Durch die Optimierung mit FastCGI müssen CGI-Prozesse nicht bei jedem Aufruf neu gestartet werden, sondern bleiben im Arbeitsspeicher erhalten und erreichen so eine Ausführungs- und Auslieferungsgeschwindigkeit, die sich nahezu mit der statischer Dateien deckt. Gleichzeitig ist die CGI-Anwendung unabhängig von einem Webserver-API programmiert.

Um dies in Greenstone nutzen zu können, muss das *Development Kit* von der FastCGI-Homepage heruntergeladen und im Verzeichnis `$GSDLHOME/packages/` unter dem Namen `fcgi`

[10] `http://www.fastcgi.com`

[11] Unter Verwaltung des W3C ⟨`http://www.w3.org/CGI/`⟩.

[12] A*pplication* P*rogramming* I*nterface*: Programmierschnittstelle, um auf die Routinen einer Anwendung von anderen Anwendungen oder Skripten aus zugreifen zu können.

```
# In Abhaengigkeit der Apache-Installation:
LoadModule fastcgi_module /usr/lib/apache/1.3/mod_fastcgi.so

# $GSDLHOME durch richtigen Pfad ersetzen

Alias /gsdl/ $GSDLHOME

<Directory $GSDLHOME/cgi-bin>
  SetHandler fastcgi-script
  Options +ExecCGI
</Directory>
```

Tafel 3.1: Konfiguration des Apache-Webservers (Auszug)

entpackt werden. In der Datei `$GSDLHOME/configure` der Greenstone-Quellcode-Distribution ist die Variable `USE_FASTCGI` auf 1 zu setzen.[13]

Nach einer Kompilierung kann das CGI-Programm über das FastCGI-Modul des Apache-Webservers betrieben werden (siehe Konfigurationsbeispiel auf Tafel 3.1).

3.4 Einrichtung

Abbildung 3.1 zeigt die Greenstone-Startseite, wie sie nach einer neuen Installation zu sehen ist. Hier finden sich hinter den Links *Greenstone* und *Documentation* Informationen zu Greenstone und weitere Links zu der Original-Dokumentation.

Standardmäßig wird die *greenstone demo*-Collection mit installiert, die als Beispiel für die Konfigurationsmöglichkeiten des Paketes dienen soll.

Die Links *The Collector* und *Administration* werden in den folgenden Abschnitten der Vollständigkeit halber erklärt, im hier vorgestellten Beispiel jedoch deaktiviert.

Eine genaue Erklärung der Einrichtung und Nutzung enthält die Original-Dokumentation mit dem *„Developer's Guide“* (Witten und Boddie 2002a) und dem *„User's Guide“* (Witten und Boddie 2002c). Die folgenden Beschreibungen bauen auf der Kenntnis dieser Dokumente auf.

3.4.1 The Collector

The Collector stellt eine grafische Oberfläche dar, um grundlegende Arbeiten an bestehenden Sammlungen durchzuführen oder neue Sammlungen einzurichten. So können auch neue Dokumente hierüber hinzugefügt werden, allerdings ohne die Möglichkeit, Metadaten manuell zu vergeben. So ist diese Oberfläche auch für Endnutzer gedacht, die ihre lokale Greenstone-Installation um weitere Dokumente erweitern wollen, ohne sich jedoch mit der näheren Funktionsweise auseinanderzusetzen.

Da sämtliche zur Pflege des Bestandes notwendigen Aktionen auf der Befehlszeile ausgeführt werden können, und hier auch die volle Flexibilität in Sachen Konfiguration gegeben ist,

[13] `USE_FASTCGI=1`

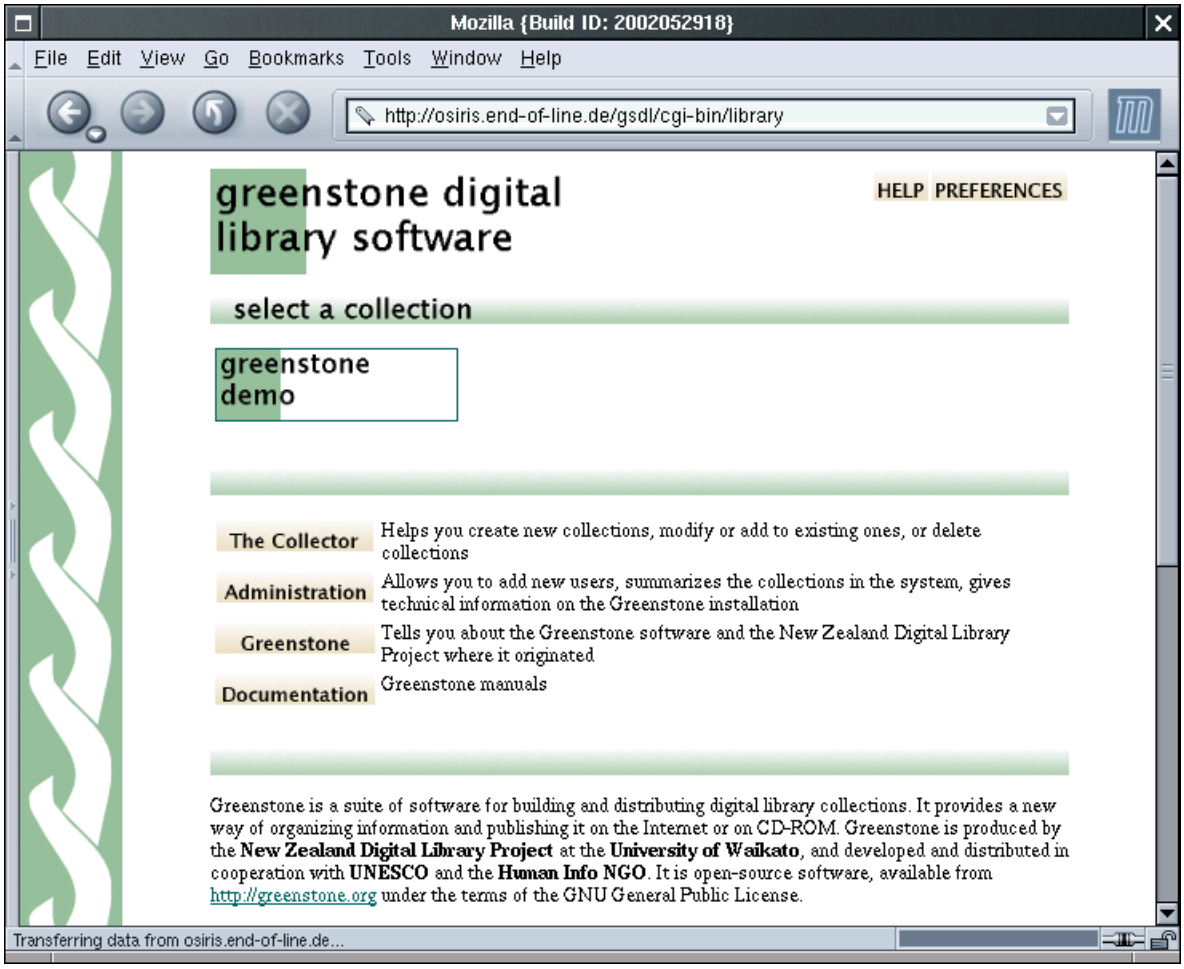

Abbildung 3.1: Startseite der Greenstone-Software nach der Installation

wird künftig allein auf diese Möglichkeit zurückgegriffen und der *Collector* für dieses Beispiel deaktiviert.

3.4.2 Administrationsoberfläche

Die *Administration* ist in ihrer Funktionalität eng verknüpft mit dem *Collector*. Neben einigen Installations- und Konfigurationsdaten, die sich hier einsehen, aber nicht ändern lassen, können hier neue Nutzer für den *Collector* angelegt und gelöscht werden.

Für die notwendigen Arbeiten ergibt sich keine weitere Funktionalität aus den gegebenen Möglichkeiten, daher wird auch die *Administration* deaktiviert.

3.4.3 Anpassung der Parameter

In der Grundinstallation sind standardmäßig schon sinnvolle Parameter vorgegeben, mit denen sich Greenstone auch gut betreiben lässt, doch lässt sich hier durch weitere Anpassungen das Paket durchaus an die eigenen Bedürfnisse angleichen.

In der Datei `$GSDLHOME/cgi-bin/gsdlsite.cfg` sind die Installationsdaten hinterlegt, auf die das CGI-Skript zurückgreifen muss. Die einzige nötige Anpassung hier ist, je nach Bedarf, der Parameter `maxrequests`, der die Lebensdauer eines CGI-Prozesses definiert, der über FastCGI optimiert wurde. Da Änderungen an der Konfiguration oder am Bestand nur bei einem Neustart der CGI-Anwendung in Kraft treten, sollte dieser Wert nicht zu hoch gewählt werden. Ein andererseits zu niedriger Wert nutzt den Optimierungsvorteil nicht genügend aus. Der optimale Wert ergibt sich aus der praktischen Erfahrung im laufenden Betrieb.

Der bedeutendste Anteil der Konfiguration findet sich in der Datei `main.cfg` im Verzeichnis `$GSDLHOME/etc/`, die alle Laufzeitparameter spezifiziert. Da die Optionen gut erklärt sind, wird hier nur auf die wichtigsten Punkte eingegangen:

`status` auf *disabled* setzen, um die *Administration* zu deaktivieren.

`collector` auf *disabled* setzen, um den *Collector* zu deaktivieren.

`macrofiles` gibt eine Liste von Makrodateien an, die sich im Verzeichnis `$GSDLHOME/macros` befinden müssen und die Erscheinung der Oberfläche definieren, wie auch die darin auftauchenden sprachspezifischen Zeichenketten. Nicht benötigte Makrodateien können hier gelöscht werden und müssen nicht geladen werden.

`Encoding ...` hier werden die Zeichensätze definiert, die für die Web-Oberfläche verfügbar sein sollen. Hier können nicht benötigte Optionen auskommentiert werden.

`Language ...` Liste der Sprachen, in denen die Web-Oberfläche verfügbar sein soll. Auch hier können nicht benötigte Optionen auskommentiert werden.

`cgiarg` definiert Standardwerte für Argumente an das CGI-Skript. Hier kann mit `cgiarg shortname=l argdefault=de` Deutsch als Standardwert für die Sprache der Web-Oberfläche gesetzt werden.

Eine minimale Beispielkonfiguration zeigt Tafel 3.2 auf der nächsten Seite

3.5 Verwaltung

Die Verwaltung von Greenstone erfolgt komplett über die Kommandozeile. Zur Vorbereitung muss im Verzeichnis `$GSDLHOME` die Umgebung vor jeder Arbeit mit dem Kommando `source setup.bash` oder `source setup.csh`, je nach verwendeter Shell, eingerichtet werden.

3.5.1 Erstellung einer Sammlung

Das Verwaltungskonzept des Bestandes bei Greenstone ist sammlungsbasiert. Das bedeutet, dass zuerst eine Sammlung, in Greenstone *Collection* genannt, angelegt werden muss, bevor Dokumente hinzugefügt werden können.

Sämtliche Einstellungen der Indizierung, etc. werden pro Sammlung angelegt.

Sämtliche Sammlungen und damit auch die variablen Daten liegen in dem Verzeichnis `$GSDLHOME/collect/`. Dort existiert für jede Sammlung ein Unterverzeichnis.

Für die eigentliche Arbeit kann das Verzeichnis der Demo-Collection gelöscht werden.

Zum Anlegen einer neuen Sammlung steht das Skript `mkcol.pl` zur Verfügung. Hier müssen die Adresse für elektronische Post und ein Kurzname für die Sammlung angegeben werden. Für dieses Beispiel sei dies die Sammlung *„diplom“*:

```
maintainer      NULL
MailServer      NULL
status          disabled
collector       disabled
logcgiargs      false
usecookies      false
LogDateFormat   LocalTime
LogEvents       disabled
EmailEvents     disabled
EmailUserEvents false

macrofiles      tip.dm style.dm base.dm query.dm help.dm pref.dm about.dm \
                document.dm browse.dm status.dm authen.dm users.dm html.dm \
                extlink.dm gsdl.dm english.dm french.dm \
                german.dm english2.dm french2.dm \
                spanish.dm spanish2.dm \
                home.dm collect.dm docs.dm \
                bsummary.dm

Encoding shortname=utf-8 "longname=Unicode (UTF-8)"
Encoding shortname=iso-8859-1 "longname=Western (ISO-8859-1)" map=8859_1.ump

Language shortname=en longname=English default_encoding=iso-8859-1
Language shortname=fr longname=French default_encoding=iso-8859-1
Language shortname=de longname=German default_encoding=iso-8859-1
Language shortname=es longname=Spanish default_encoding=iso-8859-1

pageparam       v 0

macroprecedence c,v,l

cgiarg          shortname=v longname=version multiplechar=false argdefault=0 \
                defaultstatus=weak savedarginfo=must
cgiarg          shortname=a argdefault=p
cgiarg          shortname=p argdefault=home
cgiarg          shortname=l argdefault=de
```

Tafel 3.2: Minimale Beispielkonfiguration von Greenstone in `main.cfg`

```
mkcol.pl -creator engster@iuk.hdm-stuttgart.de diplom
```

Zwar akzeptiert das Skript noch eine Reihe weiterer Optionen, die aber alle im nächsten Schritt auch über die Konfigurationsdatei der Sammlung gesteuert werden können. Da die Anpassung dieser Datei ohnehin notwendig ist, kann man die entsprechenden Parameter dabei gleich mit anpassen.

Die leere Sammlung findet sich nun mit einer eigenen Verzeichnisstruktur, auf die im Folgenden nach und nach eingegangen wird, unter `$GSDLHOME/collect/diplom/`. Die Konfiguration für die Sammlung befindet sich in der Datei `collect.cfg` im Unterordner `etc/`.

Die Bedeutung der verschiedenen Parameter ist:

`creator` nomineller Wert für den Ersteller der Sammlung, in der Regel Kontaktadresse für elektronische Post.

`maintainer` dito für den aktuellen Betreuer.

`public` Wert `true` oder `false`, der angibt, ob die Sammlung öffentlich sein soll oder nicht. Eine öffentliche Sammlung wird auf der Startseite von Greenstone aufgelistet, eine nicht-öffentliche nicht. Eine Restriktion für den Zugriff stellt diese Option jedoch nicht dar!

`indexes` eine Liste der Daten, aus denen Indizes gebildet werden sollen.

`defaultindex` Name des Indizes aus der angegebenen Liste, der in der Eingabemaske für die Suchfunktion per Voreinstellung ausgewählt sein soll.

`plugin` lädt die zum Import von verschiedenen Dateitypen notwendigen Plug-Ins.

`classify` definiert den Typ und die zu Grunde liegenden Daten für ein anzulegendes Register.

`collectionmeta` gibt weitere Steuerwerte für das Erscheinungsbild der Sammlung an.

Material

Für die hier aufzubauende Beispielsammlung soll Material verwendet werden, welches frei verteilt werden darf. Entsprechend dem Thema dieser Arbeit bieten sich hierfür, auch hinsichtlich der in den vorigen Kapiteln angeführten Quellen, relevante Dokumente aus den *Request for Comments*, des *World Wide Web Consortium* und der *Dublin Core Metadata Initiative* an. Hierbei handelt es sich um Text-, HTML- und PDF-Dateien, es werden also nur die hierfür zuständigen Plug-Ins benötigt. Dazu braucht Greenstone einige grundlegende Plug-Ins, welche interne Aufgaben erledigen.

Metadaten

Wie in Kriterium 11 auf Seite 21 gefordert, werden Metadaten nicht aus dem Dokument selbst extrahiert, sondern von einer betreuenden Person vergeben. Für dieses Beispiel wird eine Auswahl der Felder des Dublin-Core-Satzes verwendet.

Da Greenstone manche Metadaten selbst intern handhabt, wie z. B. `Source`, das den Dateinamen abbilden soll, werden weitere Metadaten mit deutlich unterschiedlichen Namen angegeben. Gemäß der Praxis bei Dublin Core geschieht dies durch das Präfix „`DC.`“:

`DC.Creator` Namen der Urheber in invertierter Form. Da Greenstone nicht mehrere Angaben des selben Metadatum kummulieren kann, werden alle Urheber in einem Feld angegeben.

`DC.Title` Titel des Dokumentes.

`DC.Source` URL des Originaldokumentes.

`DC.Date.Created` Datum der Erstellung des Dokumentes, so weit bekannt. Angabe in normierter Form: YYYY-MM-TT.[14]

`DC.Date.Available` Datum der Zugänglichmachung des Dokumentes über die digitale Bibliothek („Eingangsdatum“). Angabe ebenfalls in normierter Form.

Konfiguration

Aus den Metadaten `DC.Title` und `DC.Creator` sollen jeweils alphabetische Register gebildet werden. Für die Suche werden zwei Indizes angelegt: ein Index aus dem Dokumentenvolltext und ein Basisindex, in dem der Volltext und die Felder `DC.Title` und `DC.Creator` indexiert werden.

Hierzu müssen in der Konfigurationsdatei der Sammlung einige Anpassungen vorgenommen werden.

Die Greenstone zu Grunde liegende Retrieval-Engine existiert in zwei Versionen: das Original als `mg` und eine Neufassung als `mgpp`. Grundsätzlich bietet `mgpp` mehr Möglichkeiten als `mg`, so z. B. eine formularbasierte Suche. Die Verwendung von `mgpp` wird über die Option `buildtype` angegeben.

Zunächst werden die zu bildenden Indizes aufgelistet. Das Schlüsselwort `text` verweist auf den Dokumentenvolltext, die Metadaten können mit ihren Namen angegeben werden. Als Voreinstellung wird der Basisindex ausgewählt.

```
indexes         text,DC.Creator,DC.Title
indexes         text
defaultindex    text,DC.Creator,DC.Title
```

Anschließend werden nur jene Plug-Ins geladen, die zur Einspielung des Materials benötigt werden. Die Plug-Ins `GAPlug`, `ArcPlug` und `RecPlug` werden hierbei Greenstone-intern benötigt. Für die entsprechenden Dokumentformate werden in diesem Beispiel die Plug-Ins `TEXTPlug`, `HTMLPlug` und `PDFPlug` verwendet. Diese Plug-Ins konvertieren das einzuspielende Material in ein internes Format. Das PDF-Plug-In bietet hierbei die Möglichkeit anzugeben, ob die PDF-Dateien nach HTML oder nach reinen Text konvertiert werden sollen. Aus Effizienzgründen wird hier die Konvertierung nach reinen Text gewählt.

Das Plug-In `RecPlug` wird angewiesen, spezielle Metadaten-Dateien zu verarbeiten, aus welchen die entsprechenden Metadaten für die einzuspielenden Dokumente eingelesen werden.

```
plugin          GAPlug
plugin          TEXTPlug
plugin          HTMLPlug
plugin          PDFPlug -convert_to text
plugin          ArcPlug
plugin          RecPlug -use_metadata_files
```

[14] Die internationale Norm ISO 8601 definiert die Angabe von Datum und Uhrzeit.

Als nächstes werden über die `classify`-Anweisungen die entsprechenden Register gebildet

```
classify       AZList -metadata DC.Creator
classify       AZList -metadata DC.Title
```

Als kosmetische Einstellungen können für die Sammlung noch ein längerer Titel für die Darstellung auf den Seiten und eine Kurzbeschreibung angegeben werden. Die Einbindung von Bildern zur Repräsentation der Sammlung ist hier ebenfalls möglich. Hierbei muss darauf geachtet werden, dass sämtliche Angaben jeweils nur *innerhalb einer Zeile* stehen.

Als letztes werden den Indizes noch aussagekräftige Namen zugewiesen.

```
collectionmeta collectionname    "Beispielsammlung Diplomarbeit"
collectionmeta collectionextra   "Beispiel zur Implementierung einer
                                  digitalen Bibliothek auf Basis der
                                  <em>Greenstone Digital Library
                                  Software</em>."
collectionmeta .text,DC.Creator,DC.Title "Basisindex"
collectionmeta .text    "Volltext"
```

Die vollständige hier verwendete Konfiguration zeigt Tafel 3.3 auf der nächsten Seite.

3.5.2 Einspielen von Material

Einzuspielendes (zu „importierendes“) Material muss im Verzeichnis `import/` der entsprechenden Sammlung liegen. Hier sind auch Unterverzeichnisse erlaubt, die rekursiv importiert werden, was vor allem bei HTML-Dateien nützlich ist, da eingebundene Grafikdateien, etc. beim Speichern aus dem Browser heraus oftmals schon in Unterverzeichnissen hinterlegt werden.

Für dieses Beispiel wurde, wie oben geschildert, eine repräsentative Auswahl an themennahen Dokumenten verwendet.

Zur externen Angabe der Metadaten erwartet das `RecPlug`-Plug-In im Verzeichnis `import/` oder in einem der Unterverzeichnisse eine Datei `metadata.xml`. Die Metadaten können pro Datei oder auch für mehrere Dateien vergeben werden, wie dies Tafel 3.4 auf Seite 44 illustriert.

Das Einspielen der Dokumente übernimmt schließlich der Befehl

```
import.pl diplom
```

Damit liegt das Material jedoch lediglich im Greenstone-internen Format im Verzeichnis `archives/` vor. Beim Import wurden aus den Dateien MD5-Summen gebildet, die fortan als Identifikator der Dokumente fungieren.

Bei einem erneuten Einspielen desselben Materials werden die vorhandenen Kopien *nicht* ersetzt. Statt dessen finden sich nach dem Import jeweils zwei Kopien davon wieder. Die Option `-removeold` sorgt dafür, dass *alle* bisherigen Archiv-Inhalte vor dem Einspielen gelöscht werden, was dann nützlich ist, wenn alle Materialien neu importiert werden.

```
import.pl -removeold diplom
```

Damit die zur Nutzung über das CGI-Interface notwendigen Daten-Dateien[15] vorliegen, müssen aus diesen Archiv-Dateien die Indizes und Register gebildet werden. Dies erledigt der Befehl

[15] Datenbank-ähnliche Dateien mit den Indizes und Informationen zur *Collection*.

```
creator       engster@iuk.hdm-stuttgart.de
maintainer    engster@iuk.hdm-stuttgart.de
public        true
buildtype     mgpp

indexes        text,DC.Creator,DC.Title
indexes        text
defaultindex   text,DC.Creator,DC.Title

plugin         GAPlug
plugin         TEXTPlug
plugin         HTMLPlug
plugin         PDFPlug -convert_to text
plugin         ArcPlug
plugin         RecPlug -use_metadata_files


classify       AZList -metadata DC.Creator
classify       AZList -metadata DC.Title

collectionmeta collectionname    "Beispielsammlung Diplomarbeit"
# Zeilenumbruch nur für den Ausdruck!
collectionmeta collectionextra   "Beispiel zur Implementierung einer
                                  digitalen Bibliothek auf Basis der
                                  <em>Greenstone Digital Library
                                  Software</em>."
collectionmeta .text,DC.Creator,DC.Title "Basisindex"
collectionmeta .text    "Volltext"
```

Tafel 3.3: Konfiguration der Sammlung (`collect.cfg`)

```
<?xml version="1.0" encoding="iso-8859-1" standalone="no"?>
<!DOCTYPE DirectoryMetadata SYSTEM
 "http://greenstone.org/dtd/DirectoryMetadata/1.0/DirectoryMetadata.dtd">
<DirectoryMetadata>

  <FileSet>
    <FileName>.*</FileName>
    <Description>
      <Metadata name="DC.Date.Available">
        2002-09-18
      </Metadata>
    </Description>
  </FileSet>

  <FileSet>
    <FileName>rfc1321.txt</FileName>
    <Description>
      <Metadata name="DC.Creator">
        Rivest, Ronald L. ; RSA Data Security
      </Metadata>
      <Metadata name="DC.Title">
        The MD5 Message-Digest Algorithm
      </Metadata>
      <Metadata name="DC.Source">
        http://www.ietf.org/rfc/rfc1321.txt
      </Metadata>
      <Metadata name="DC.Date.Created">
        1992-04
      </Metadata>
    </Description>
  </FileSet>

</DirectoryMetadata>
```

Tafel 3.4: Angabe der Metadaten in der Datei `metadata.xml`

```
buildcol.pl diplom
```

Nach dessen Durchlauf liegen sämtliche Daten-Dateien im Verzeichnis `building/` vor. Der teilweise recht lange Prozess eines *build* verläuft daher unabhängig von bereits bestehenden Daten-Dateien, die für den produktiven Betrieb im Verzeichnis `index/` liegen.

Zur Aktualisierung müssen die alten Dateien durch die neuen ersetzt werden:

```
rm -fr index/* ; mv building/* index/
```

Ist das CGI-Programm mit FastCGI optimiert, so werden die neuen Daten-Dateien erst beim Neustart des Prozesses eingelesen, was je nach angegebener Lebensdauer auch recht lange dauern kann. Ein sanfter Neustart des Webservers führt zum sofortigen erneuten Einlesen *aller* benötigten Daten-, Konfigurations- und Makrodateien.

3.6 Nutzung

Nachdem die neu erstellte Sammlung für die Greenstone-Software verfügbar ist, taucht sie auch in der Liste der Sammlungen auf der Startseite auf (siehe Abbildung 3.2), sofern nicht die Option `public false` angegeben wurde.

Abbildung 3.2: Liste der verfügbaren Sammlungen auf der Startseite

3.6.1 Darstellung

Zwar ist als Sprache „Deutsch“ gewählt, doch hängt das Erscheinungsbild der Oberfläche auch davon ab, in welchem Maße Übersetzungen verfügbar sind. So tauchen in der aktuellen Greenstone-Version immer wieder Vermischungen der originalen englischen Gestaltung und der nur teilweise vorgenommenen deutschen Übersetzung auf.

Durch Anpassung der Makrodateien kann dies allerdings behoben oder den eigenen Wünschen gemäß gestaltet werden. Ein Vorteil freier Software ist hier, dass eine vorgenommene vollständige Übersetzung Einzug in die ständige Weiterentwicklung der Software halten kann und somit allen weiteren Benutzern zur Verfügung steht.

Dem entsprechenden Eintrag folgend gelangt man zur Hauptseite der Sammlung (siehe Abbildung 3.3 auf der nächsten Seite), die alle hierfür verfügbaren Funktionalitäten anbietet sowie die in der `collect.cfg` angegebene Kurzbeschreibung darstellt und knappe Hinweise zur Benutzung der verschiedenen Optionen wie der Suche oder den Registern gibt.

Diese Optionen stehen in der Zeile unter dem Titel der Sammlung zur Verfügung. Hinter der Schaltfläche *Suche* steht der Zugang zu der Suchseite, die sich je nach Einstellung von dem auf dieser ersten Seite angebotenen Eingabefeld zur Suche unterscheidet. Auf die Funktion der Suche wird in Abschnitt 3.6.3 auf Seite 49 näher eingegangen.

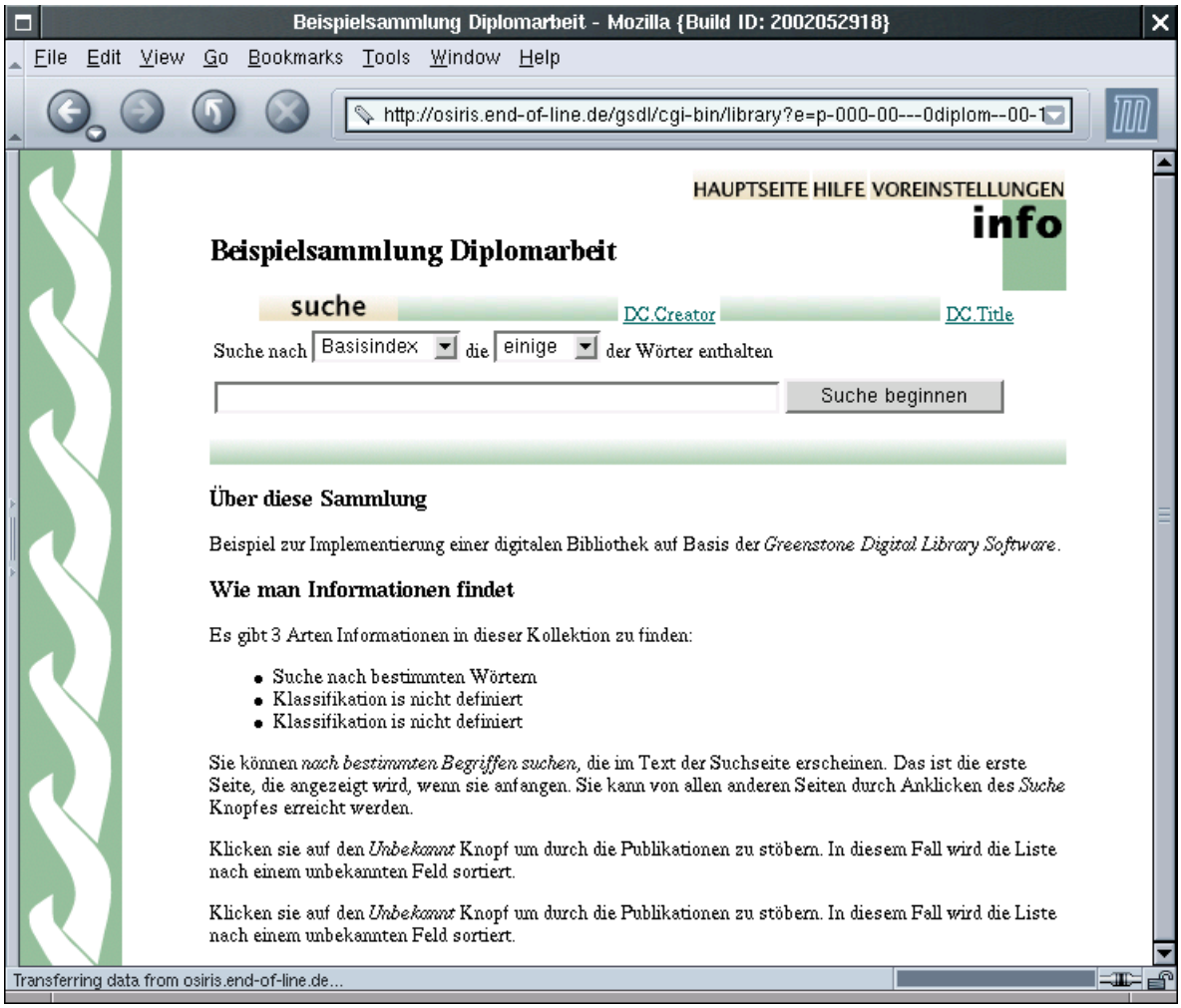

Abbildung 3.3: Hauptseite der Sammlung „*diplom*“

Für das Autoren- und Titelregister stehen auffälliger Weise die Namen der dafür verwendeten Dublin-Core-Felder an der Stelle der Schaltflächen. Dies hängt damit zusammen, dass Greenstone für diese Metadatennamen automatisch keine Schaltflächen vergeben kann. Durch eine Anpassung in der `collect.cfg` können hierfür die bereits vorhandenen Schaltflächen verwendet werden. Diese Anpassungen sind nach einem neuen *build* der Sammlung verfügbar.

```
classify       AZList -metadata DC.Creator -buttonname Creator
classify       AZList -metadata DC.Title -buttonname Title
```

Das Ergebnis ist in Abbildung 3.4 zu sehen.

Abbildung 3.4: Schaltflächenleiste nach der Anpassung

Bei der Anwahl des Autorenregisters, wie auch des Titelregisters, erscheint die Liste in einer ernüchternden Darstellung. Angezeigt werden die von Greenstone selbst extrahierten Metadaten, in beiden Fällen der zu ermitteln versuchte Titel, was an sich teils brauchbare Ergebnisse aufweist, aber kaum als genügend angesehen werden kann (vgl. Abbildung 3.5).

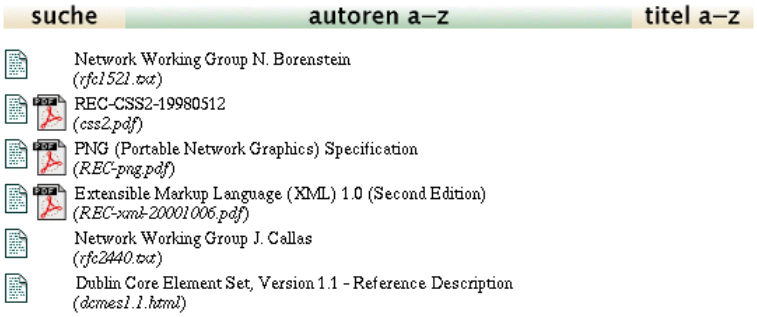

Abbildung 3.5: Autorenregister ohne Anpassung

Die Darstellung der Titellisten der Register, wie auch der Liste der Suchergebnisse, kann durch entsprechende Konfigurationen in der `collect.cfg` angepasst werden.

Die Einträge in der Liste erfolgen als Tabellenreihe. Auf die Metadaten kann durch spezielle Platzhalter zugegriffen werden. Links werden angegeben zu dem Greenstone-internen Format und zu den Quelldateien, soweit diese vorhanden sind, wie in diesem Beispiel die PDF-Dateien. Beim Zugriff auf HTML- und Textdateien steht nur das Greenstone-interne Dokument zur Verfügung.

Bei der Konfiguration der Darstellung ist wieder auf eine *einzeilige Angabe* zu achten. Die entsprechenden Register werden über den Schlüssel `CL*VList` angesprochen, wobei der Stern für die Nummer des Registers steht. Da das Autorenregister als erstes in der Konfigurationsdatei genannt ist, trägt dieses auch die Nummer 1. Die Ergebnisliste der Suchfunktion wird über `SearchVList` angegeben. In der in Tafel 3.5 auf der nächsten Seite angegebenen Konfiguration unterscheiden sich Titel- und Autorenregister nur durch die Hervorhebung der jeweils interessanten Metadatenfelder. Die Suchergebnisliste ist dem Autorenregister gleich. Das Ergebnis ist in Abbildung 3.6 auf Seite 49 zu sehen.

```
# mehrzeilige Darstellung nur für den Ausdruck!
format CL1VList "<td>
                   [link]_icontext_[/link][srclink][srcicon][/srclink]
                 </td>
                 <td>
                   <strong>[DC.Creator]:</strong> [DC.Title]. -
                   [DC.Date.Created] (Zugang: [DC.Date.Available])<br>
                   Quelle:
                   <samp><a href='[DC.Source]'>[DC.Source]</a></samp><br>
                   <hr noshade>
                 </td>"

format CL2VList "<td>
                   [link]_icontext_[/link][srclink][srcicon][/srclink]
                 </td>
                 <td>
                   [DC.Creator]: <strong>[DC.Title]</strong>. -
                   [DC.Date.Created] (Zugang: [DC.Date.Available])<br>
                   Quelle:
                   <samp><a href='[DC.Source]'>[DC.Source]</a></samp><br>
                   <hr noshade>
                 </td>"

format SearchVList "<td>
                      [link]_icontext_[/link][srclink][srcicon][/srclink]
                    </td>
                    <td>
                      <strong>[DC.Creator]:</strong> [DC.Title]. -
                     [DC.Date.Created] (Zugang: [DC.Date.Available])<br>
                     Quelle:
                     <samp><a href='[DC.Source]'>[DC.Source]</a></samp><br>
                     <hr noshade>
                   </td>"
```

Tafel 3.5: Anpassung der Listendarstellung in `collect.cfg`

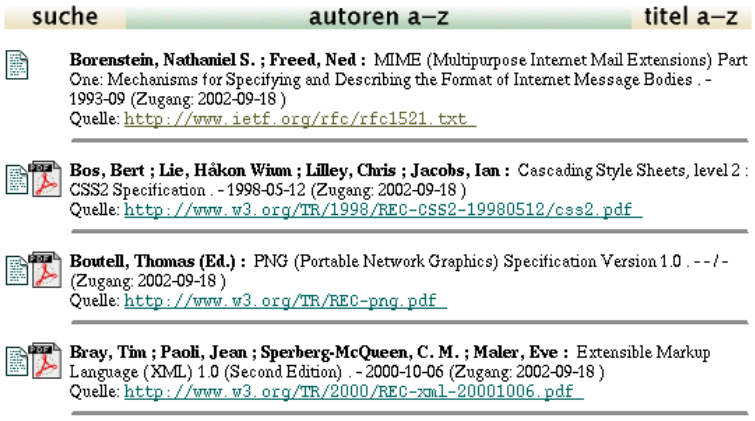

Abbildung 3.6: Autorenregister nach der Anpassung

3.6.2 Benutzereinstellungen

Über die Schaltfläche *Voreinstellungen* kann eine Seite mit Optionen für den Zugriff auf die Sammlung aufgerufen werden.

Für das Erscheinungsbild der Oberfläche können hier die Schnittstellensprache, der Schriftsatz, sowie das Erscheinungsbild ausgewählt werden. Letzteres kann grafisch gewählt werden, was per Voreinstellung der Fall ist, oder textuell. Im textuellen Modus werden keine Grafiken geladen, wodurch der Seitenaufbau beschleunigt wird.

Hier sind auch zahlreiche Optionen für die Suchfunktion verfügbar, die in Abschnitt 3.6.3 behandelt werden.

3.6.3 Retrieval

Auf der Seite der *Voreinstellungen* können, wie erwähnt, die Optionen für die Suchfunktion festgelegt werden. Am effektivsten ist hier die *form search* beim *Type of search*. Nach deren Auswahl steht eine geänderte Liste an Optionen zur Verfügung. So ist es hier sinnvoll, als *Form type* die Option *advanced* zu wählen, bei der sich für laufende Suchen die wichtigsten Einstellungen während der Arbeit ändern lassen. Die ausgewählten Optionen werden mit der Schaltfläche *Update settings* übernommen.

Grundsätzlich steht ein Eingabefeld für die Suche auf der Startseite der Sammlung zur Verfügung (wie in Abbildung 3.3 auf Seite 46 zu sehen ist). Das vollständige Suchfeld erhält man über die Schaltfläche *Suche* (abgebildet auf der nächsten Seite).

Hier bieten sich einige Auswahlmöglichkeiten: in den oberen Auswahlfeldern kann man zwischen den zu durchsuchenden Indizes wählen, in diesem Fall zwischen dem Basisindex und dem Volltextindex. Eine weitere Option erlaubt die Angabe der Ordnung der Ergebnisse: ob in gestaffelter Ordnung (*ranked*) oder nach der physikalischen Ordnung in den Daten-Dateien (*natural*).

Die hinter den Eingabefeldern stehenden Auswahllisten erlauben dann die genaue Spezifizierung, in welchem Feld *innerhalb* des ausgewählten Index der Suchbegriff vorkommen muss. Hierbei fehlt es an einer Abbildung der Greenstone-internen Feldnamen auf normalverständliche Namen. Die Bedeutung dieser Schlüssel beschreibt Tafel 3.6 auf Seite 51, wobei sich dies je nach Benennung der Felder unterscheidet. Leider taucht diese Auswahlmöglichkeit

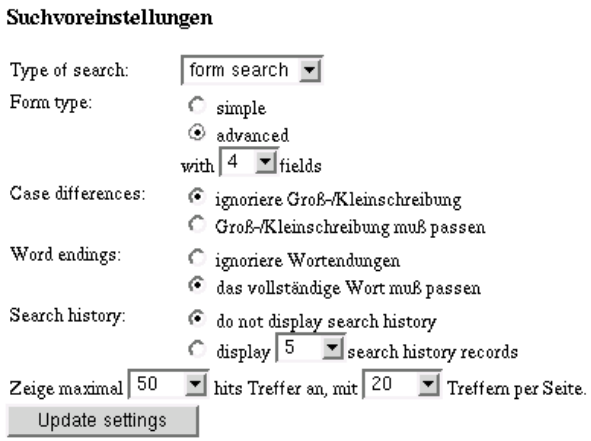

Abbildung 3.7: Optionen für die Suchfunktion

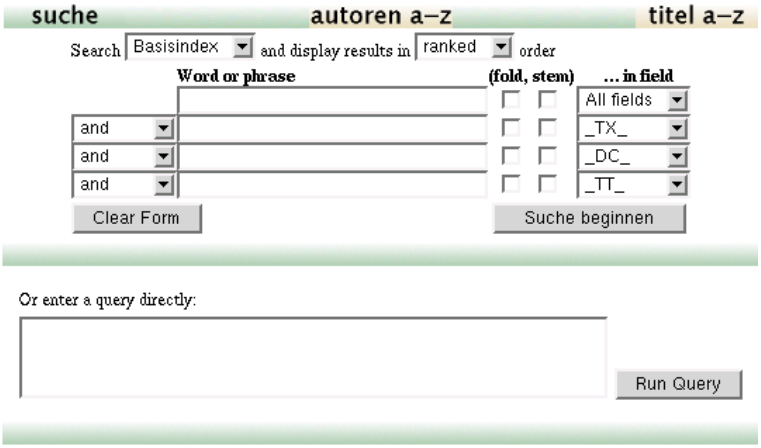

Abbildung 3.8: Vollständiges Suchformular im *advanced*-Modus.

grundsätzlich auf, egal ob die genannten Felder im ausgewählten Index vorhanden sind oder nicht.

Die vor den Eingabefeldern stehenden Operatoren entsprechen den grundlegenden boole'schen Operatoren. Hier wird also eine boole'sche Feldsuche mit einer gestaffelten Ergebnismenge verknüpft.

Die Optionen *fold* und *stem* erlauben eine Vorverarbeitung des Suchbegriffes: *fold* ignoriert Unterschiede in der Groß- und Kleinschreibung, *stem* versucht durch einen Algorithmus den Wortstamm zu ermitteln und nach diesem in einem eigens dafür aus den Wortstämmen gebildeten Index zu suchen, der während dem *build* der Sammlung parallel angelegt wird. Zu jedem vom Verwalter definierten Index existiert ein entsprechender Index aus den Wortstämmen.

Auffallend ist, dass keine Trunkierung, also eine Suche nach nur teilweise angegebenen Begriffen, möglich ist. Dies erklärt sich dadurch, dass die Suche in Greenstone nach dem *Dictionary*-Prinzip arbeitet: der (gegebenenfalls vorverarbeitete) Suchbegriff wird mit den im Index vorkommenden Begriffen abgeglichen und auf 100%ige Übereinstimmung hin überprüft.

Schlüssel	Bedeutung
All fields	in beliebigen Feldern des Index
_TX_	*nur* im Dokument-Volltext
_DC_	*nur* im Feld `DC.Creator`
_TT_	*nur* im Feld `DC.Title`

Tafel 3.6: Bedeutung der Schlüssel für die Metadatenfelder

Dieses Prinzip ermöglicht eine äußerst schnelle Verarbeitung der Suchanfrage, allerdings auf Kosten einer nicht möglichen Trunkierung. Vergleichbar wäre dies mit der Suche in einem Wörterbuch, bei dem nur ganze vorgegebene Wörter gesucht werden könnten.

Das Textfeld unter dem Suchformular erlaubt die Eingabe einer komplexen Suchanfrage mit der Syntax der zu Grunde liegenden Retrieval-Engine `mgpp`. Die im Formular angegebene Suchanfrage wird hier automatisch „übersetzt“ und eingetragen.

Abbildung 3.9 auf Seite 54 zeigt das Ergebnis einer so erfolgten Suche.

3.6.4 Geführte Suche mit PHIND

PHIND (für P*hrase* H*ierarchy* IND*ex*) ist in seinen Funktionen und Grundlagen ausführlich bei Paynter und Witten (2001) beschrieben.

Hinter PHIND steht ein Browser für Schlüsselsätze, der sich mit einem Thesaurus kombinieren lässt. Leider ist letzteres Verfahren noch nicht dokumentiert, ein funktionstüchtiges Beispiel hierfür findet sich jedoch auf den Seiten der *New Zealand Digital Library* in der Sammlung *FAO on the Internet (1998)*[16].

PHIND implementiert eine *keyword-in-context*-ähnliche Suche, die auf aus den Dokumenten extrahierten Kernsätzen beruht. Der Gedanke dahinter ist, den Nutzer bei seiner Suche nach einzelnen Begriffen dadurch zu unterstützen, dass die Zusammenhänge aufgezeigt werden, in denen diese Begriffe stehen.

Um PHIND nutzen zu können, muss dies als *classifier* in der `collect.cfg` eingetragen werden.

```
classify phind -text document:text
```

Dabei wird aus dem Dokumententext der Index gebildet, zu dessen Erstellung auch andere Felder, z. B. Kurzreferate oder ähnliches, verwendet werden können. Jedoch ist dies wenig sinnvoll, da bei der Verwendung des Volltextes der größte Nutzen erzielt werden kann. Allerdings ist dessen Aufbau in Abhängigkeit der zu bearbeitenden Textmasse sehr rechen- und zeitaufwändig.

Abbildung 3.10 auf Seite 55 zeigt die Recherche mit dem PHIND-Browser, der als Java-Applet realisiert ist. Per Voreinstellung werden als Dokumentitel (blau) die von Greenstone extrahierten *Title*-Daten angezeigt. Dies kann über die *classify*-Anweisung gesteuert werden.

Im Beispiel wird nach dem Wort *„core“* gesucht, das weiterführen kann zu *„Dublin Core“*, und dies wiederum weiter zu *„Dublin Core Metadata“*. Die darauf zutreffenden Dokumente schränken sich dabei immer weiter ein.

[16] `http://www.nzdl.org/cgi-bin/library?a=p&p=about&c=fi1998`

Bei einer Auswahl eines Dokumenttitels öffnet sich ein neues Browser-Fenster, in dem das Dokument im Greenstone-internen Format angezeigt wird.

3.6.5 Dokumentansicht

Ohne weitere Einstellung stellt sich die Ansicht eines Dokumentes wie in Abbildung 3.11 auf Seite 56 dar. Dort sind auch vereinzelt auftretende Probleme mit Sonderzeichen zu sehen. Das Aussehen von HTML-Dateien wird allerdings so gut als möglich originalgetreu wiedergegeben und die zuletzt eingegebenen Suchbegriffe farblich hervorgehoben.

Der unter der Schaltflächenleiste auftauchende Titel soll hier um die vollständige Angabe der Metadaten erweitert werden. Die geschieht über eine `format`-Anweisung in der `collect.cfg`, die ebenfalls wieder *einzeilig* erfolgen muss:

```
format   DocumentHeading   "<p><strong>[DC.Creator]:</strong> 
   [DC.Title]. - [DC.Date.Created] (Zugang: [DC.Date.Available])<br>
   Quelle: <samp><a href='[DC.Source]'>[DC.Source]</a></samp><br></p>
   <p>[srclink][srcicon][/srclink]</p><hr noshade>"
```

Existiert zu dem Dokument eine Quelldatei, z. B. das originale PDF-Dokument, wird zusätzlich ein Link dazu eingefügt. Das Ergebnis der Anpassung zeigt Abbildung 3.12 auf Seite 56 für das Dokument über CSS, Level 2.

3.6.6 Interoperabilität

Bainbridge u. a. (2001) gehen anschaulich auf die Möglichkeiten der Interoperabilität bei Greenstone ein. Die am weitesten fortgeschrittene und brauchbarste Schnittstelle ist die zu CORBA-Clients.

Leider ist das CORBA-Protokoll sehr umständlich und komplex und daher auch nicht leicht zu implementieren, bietet dafür jedoch weitreichende Möglichkeiten. So können verschiedene Greenstone-Server wie *eine einzige* Anwendung fungieren, deren Komponenten aber auch über Netzwerke verteilt sein dürfen.

Leider mangelt es bei Greenstone an einfachen und inzwischen weit verbreiteten Schnittstellen wie etwa dem OAI-Protokoll. Jedoch ist es möglich und denkbar, auf Basis der CORBA-Schnittstelle einen OAI-Zugang oder eine Web-Services-Schnittstelle aufzusetzen.

Dies gilt natürlich auch für den Fall, in dem ein Greenstone-Server als Client für andere Dienste fungiert.

Zur näheren Darstellung sei auf oben genannten Titel verwiesen.

3.7 Distribution auf CD-ROM

Eine Besonderheit bei Greenstone ist die Möglichkeit, Sammlungen zu „exportieren“ und auf CD-ROM zu verteilen. Auf diese Weise erstellte CD-ROMs können auf Windows-PCs installiert werden und stellen dort die nahezu gleiche Umgebung wie über den Serverdienst zur Verfügung. Ausnahme bildet hier das PHIND-Modul, dessen Export auf CD-ROM zur Zeit noch nicht funktioniert.

Bei der Installation kann gewählt werden, ob die Daten-Dateien auf der CD-ROM bleiben oder lokal auf die Festplatte kopiert werden sollen. Je nach Größe dieser Dateien muss hier eine Entscheidung zwischen Speicherplatzbelegung und Zugriffsgeschwindigkeit gefällt werden.

Zur Nutzung der CD-ROM startet die Greenstone-Software aus dem Startmenü heraus einen lokalen Server und einen Web-Browser, der ausgewählt werden kann. Über ein Netzwerk könnte dieser Dienst auch anderen Teilnehmern zur Verfügung stehen.

Zur Distribution auf CD-ROM muss die entsprechende Sammlung zunächst „exportiert“ werden. Diese Aufgabe erledigt der Befehl

```
exportcol.pl diplom
```

Damit dies funktionieren kann, muss das separate Paket `gsdl-2.35-export.zip` zuvor im Verzeichnis `$GSDLHOME/bin/windows/` ausgepackt werden.

Nach einem erfolgreichen Export liegen die notwendigen Dateien im Verzeichnis `$GSDLHOME/tmp/exported_diplom/` vor. Dieser Verzeichnisbaum muss anschließend nur noch auf CD-ROM gespeichert werden.

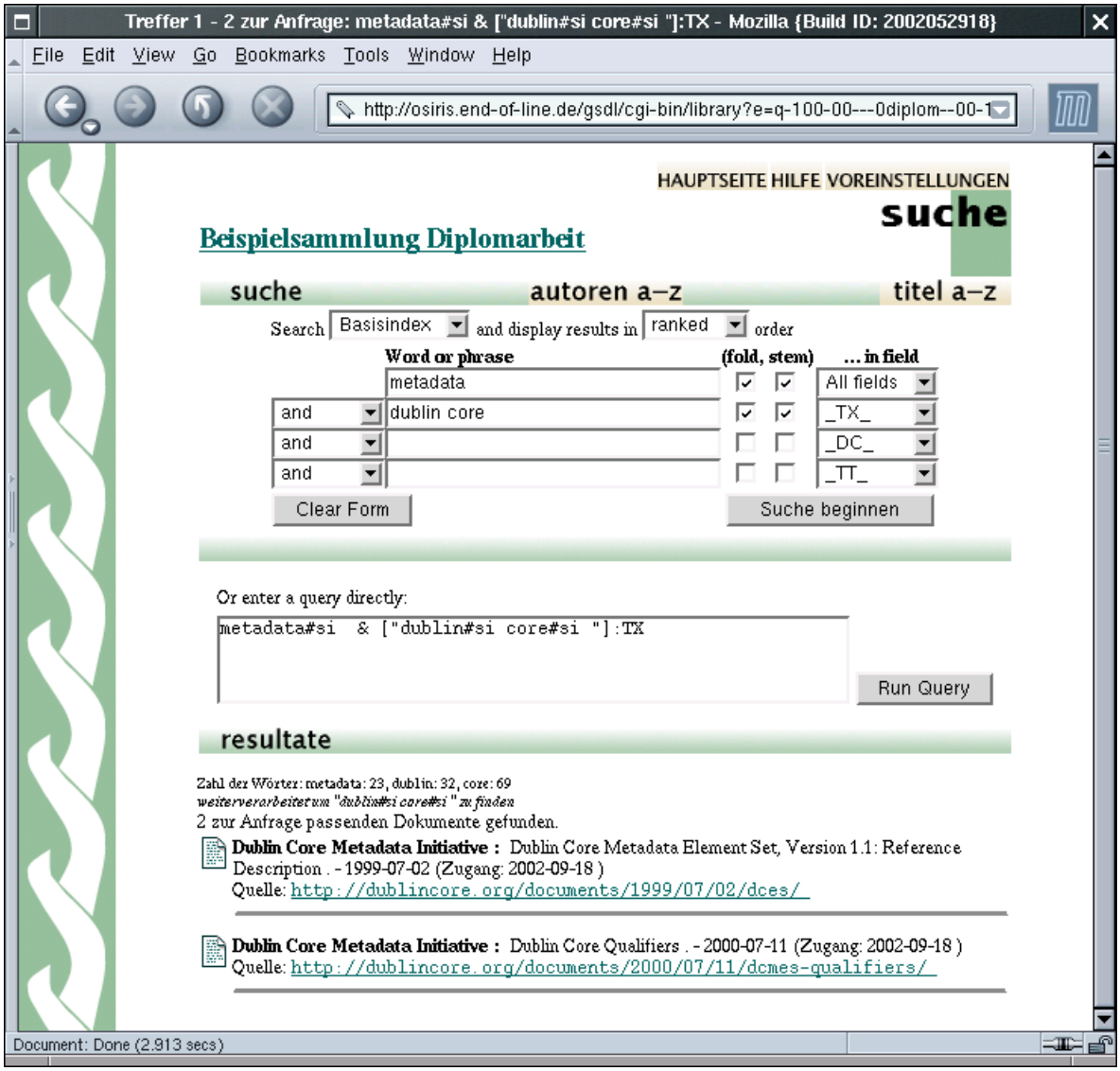

Abbildung 3.9: Suchergebnis bei der fortgeschrittenen Formularsuche

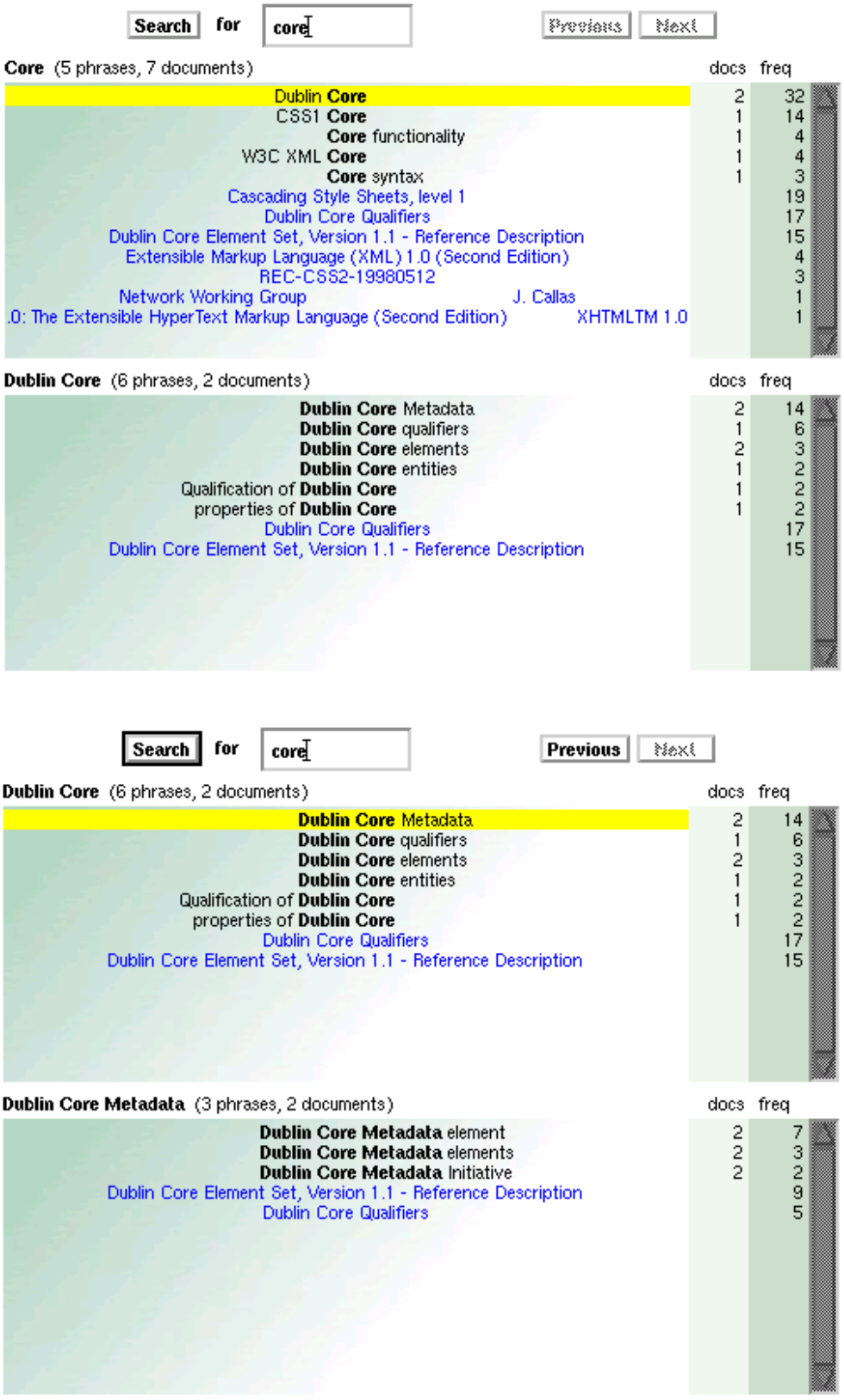

Abbildung 3.10: Beispielsuche mit PHIND

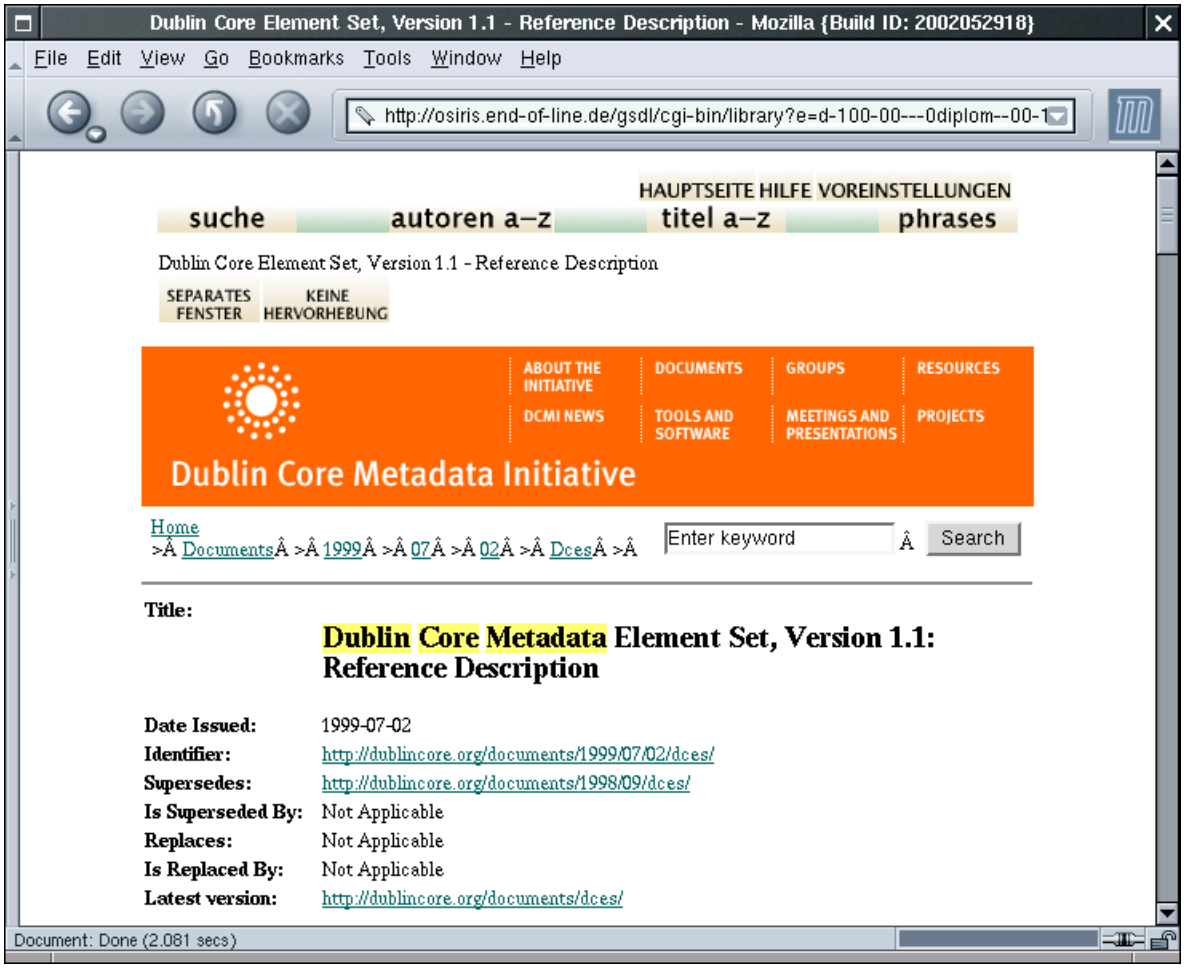

Abbildung 3.11: Dokumentenansicht in Greenstone

Abbildung 3.12: Anpassung des Dokument-Titels

Kapitel 4

Resumée

4.1 Zur Implementierung

Die versuchte Umsetzung der in Kapitel 2 formulierten Kriterien kann durch die *Greenstone Digital Library Software* bei weitem nicht geleistet werden. So kann das Paket zwar mit für diesen Zweck gelungenen Techniken und Konzepten aufwarten, überzeugt bei deren endgültiger Umsetzung für den Produktivbetrieb aber eher weniger.

Hier macht sich der akademische Charakter des Greenstone-Projektes negativ bemerkbar: dessen Ziel ist es in erster Linie softwaregestützte Techniken zu entwickeln, die für den Einsatz in digitalen Bibliotheken brauchbar sind. Zur vollständigen Umsetzung eines Paketes, das auch den in dieser Arbeit formulierten Kriterien genügt, fehlen hier genau jene konzeptionellen Hintergründe aus bibliothekarischer Sicht.

Ist von der Installation und Einrichtung der Software einmal abzusehen, so fehlt Greenstone eine endnutzerorientierte Oberfläche, welche die Verwaltung sämtlicher Sammlungen und Dokumente uneingeschränkt erlaubt. Die lediglich grundlegendsten Aufgaben können mit dem *Collector* erledigt werden, doch sämtliche weiterführende Arbeiten wie Korrektur der Metadaten, Löschen von Material oder das Anlegen interner Bezüge (Verknüpfungen, *trails*) sind entweder nur durch Handarbeit auf Ebene des Host-Systems durch direkte Bearbeitung des Materials oder gar nicht möglich. Was man sich hier für Sammlungen geringen Umfangs oder für einmalige Projekte noch vorstellen könnte, ist bei wachsender Komplexität des Bestandes zum Scheitern verurteilt.

Bei der Frage der Interoperabilität erscheint Greenstone ebenfalls wieder zu sehr akademisch. So setzt das Entwicklerteam auf die im großen Maßstab sehr sinnvoll erscheinende Schnittstelle für CORBA, lässt aber einfach zu implementierende und inzwischen bewährte Standards wie das OAI-Protokoll unbeachtet. Hier kann zu Recht darauf verwiesen werden, dass dessen Einbindung über die Schnittstellen-Architektur durchaus zu leisten ist, jedoch darf man die Verfügbarkeit dieser Schnittstelle in der Standardvariante der Software erwarten.

Vorteilhaft sind die geringen Systemanforderungen der Software, die allerdings durch einige wenige Abstriche in der Funktionalität erreicht werden, wie etwa die fehlende Möglichkeit zur trunkierten Suche. Hier liegt das Argument nahe, dass angesichts der immer weiter steigenden Leistungsfähigkeit von Servern und PCs diese Überlegungen nur zweitrangig sind, jedoch gehen Witten u. a. (1999b, S. 460 ff.) auf Sammlungen mit einer Größe von mehreren Gigabytes ein, welche die Notwendigkeit derartiger Optimierungsmaßnahmen einleuchtend veranschaulichen. Dass es nicht einmal der keineswegs an mangelndem Interesse leidende Internet-Suchdienst

Google für notwendig erachtet, eine Trunkierung der Suche (wohl aus eben den gleichen Gründen) anzubieten, zeigt, dass diese Option zwar sinnvoll aber dennoch keineswegs unbedingt notwendig ist.

Zusätzliche Möglichkeiten wie PHIND oder PHRASIER, deren experimentelle Natur immer noch deutlich zu sehen ist, geben trotz der Mängel beim Einsatz zahlreiche Anregungen für erweiterte Suchmethoden, mit denen eine sinnvolle Erschließung des Bestandes ohne allzu großen Aufwand möglich ist.

Abschließend kann der *Greenstone Digital Library Software* zwar ein hohes Entwicklungspotential bescheinigt werden, das aber durch eine immer wieder mangelnde oder bruchstückhafte Umsetzung der möglichen und vorhandenen Techniken ausgebremst wird.

4.2 Zur Konzeption

Die formulierten Kriterien für digitale Bibliotheken sind sehr ausschließlich und umfassen einen sehr eng umgrenzten Bereich. In der Tat werden die wenigsten oder gar keine der bisherigen *digitalen Bibliotheken* allen diesen Kriterien entsprechen.

Dies muss aber auch nicht unbedingt sein. Sinn einer Konzeption ist es, ein möglichst universelles Rahmenwerk zu formulieren. Die in Kapitel 2 formulierten Kriterien stehen hierbei vor allem unter den Bedingungen der freien Information. Konkrete Umsetzungen sind aber immer spezifisch und weisen deutlich mehr Faktoren auf, als eine allgemeine Konzeption berücksichtigen kann.

Der vorliegende Kriterienkatalog umfasst die wichtigsten Aspekte digitaler Bibliotheken und geht dabei auf die relevanten Themen ein. Er soll daher bei konkreten Umsetzungen als Ratgeber dienen.

Teil III

Anhang

Anhang A

Dokumentformate

A.1 SGML/XML

Die als ISO-Norm vorliegende *Standard Generalized Markup Language* (ISO 1986) ging aus der 1969 bei IBM von Charles Goldfarb, Edward Mosher und Raymond Lorie entwickelten GML

Mit (S)GML wurde erstmals das Prinzip des zwei Jahre zuvor von William Tunnicliffe beschriebenen *„generic coding“* umgesetzt, das wiederum auf den Buch-Designer und Setzer Stanley Rice zurückgeht.

Die Idee dahinter ist, die Struktur eines Textes unabhängig von seiner grafischen Darstellung zu beschreiben, ihm also Attribute wie „Überschrift“, „Zitat“ und dergleichen mitzugeben anstatt „16 Punkt, fett“. Die Frage der Darstellung sollte sich so unabhängig vom eigentlichen Dokument klären lassen.

SGML ist eine Metasprache, mit der *generic coding* erstmals im großen Einsatz möglich war. Mit ihrer Hilfe können Sätze von Markierungen (*„tags“*) definiert werden, um die der Inhalt einer Textdatei erweitert werden kann. Hintergrund ist vor allem die maschinelle Verarbeitung von Text, da die Auszeichnung mit Tags der verarbeitenden Anwendung Informationen über die *Funktion* des folgenden Inhaltes gibt. Als bekannteste SGML-Andwendung gilt HTML.[1]

Da SGML auf Grund seiner Struktur teils recht kompliziert werden kann, bestand von Anfang an der Bedarf nach einer einfacheren Untermenge. Dieser Forderung wurde mit der Entwicklung der *eXtensible Markup Language*[2] (XML) entsprochen.

Eine sehr gute Behandlung der Möglichkeiten von SGML und XML hinsichtlich ihrer Bedeutung für die Kodierung von Information findet sich bei Lobin (2000).

Durch ihre Basierung auf Textdateien sind SGML/XML-Dateien auf allen Systemen einfach hand zu haben. Gleichzeitig erlaubt ihre Strukturierung eine logische Verarbeitung der Inhalte.

Durch Stylesheets[3] ist es möglich, die Darstellung einer solchen Datei oder deren Umwandlung in ein anderes Format zu steuern. Zudem sind auch verschiedene *views* möglich, bei denen ein und dieselbe Datei verschiedenartig in Erscheinung tritt, indem z. B. nur die

[1] *„HyperText Markup Language“* ⟨`http://www.w3.org/MarkUp/`⟩

[2] `http://www.w3.org/XML/`

[3] „Stilvorlagen“, welche die Darstellungsinformationen zu strukturierten Dokumenten separat definieren, z. B. CSS ⟨`http://www.w3.org/Style/CSS/`⟩, XSL und XSLT ⟨`http://www.w3.org/Style/XSL/`⟩ und DSSSL ⟨`http://www.jclark.com/dsssl/`⟩.

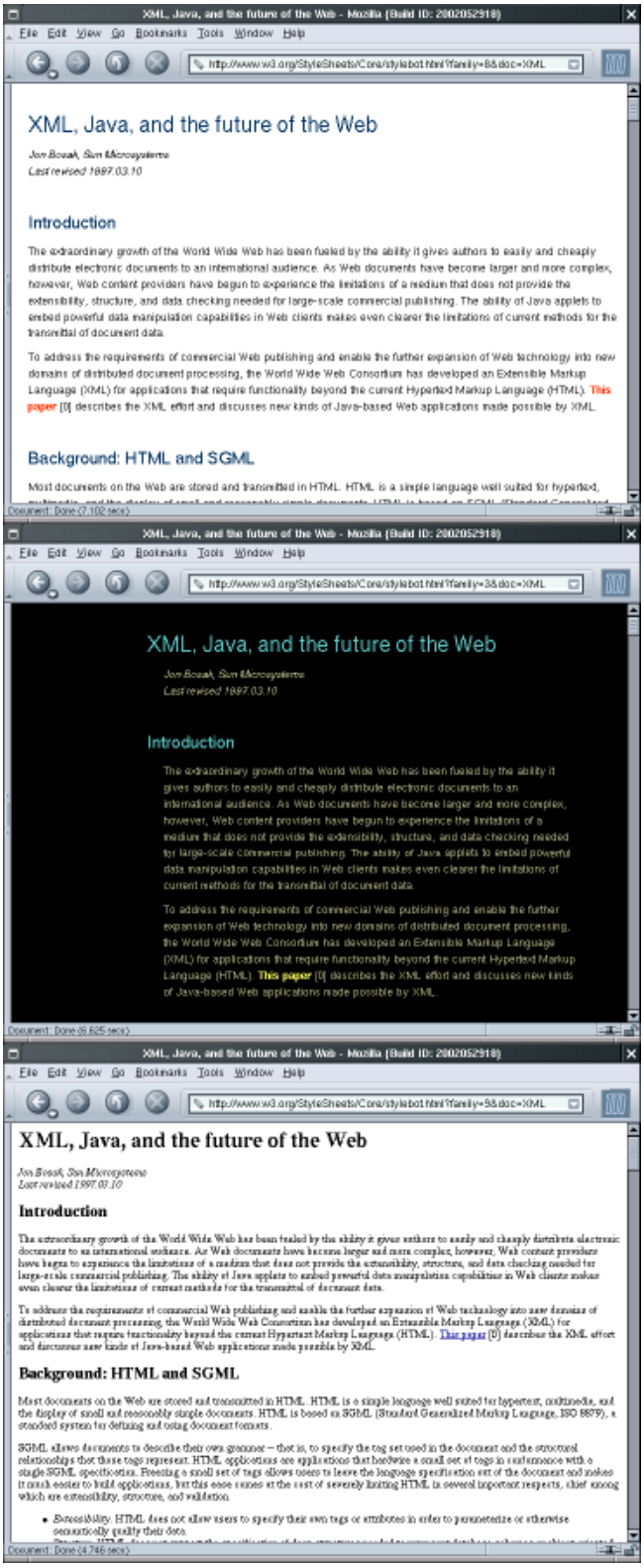

Abbildung A.1: Ansicht derselben Datei mit verschiedenen Stylesheets

Überschriften oder nur der Textkörper, aber keine Abbildung angezeigt werden, was nicht nur die Verarbeitung für verschiedene Medien und Ausgabeformate vereinfacht, sondern auch die gezielte Suche und Extraktion von gespeicherten Daten oder Information.

Abbildung A.1 zeigt die Darstellung ein und derselben Seite mit zwei unterschiedlichen und keinem Stylesheet und ist dem Beispiel des W3C entnommen[4].

A.2 PostScript und PDF

```
%!PS-Adobe-2.0 EPSF-2.0
%%Title: eps-Beispiel
%%BoundingBox: 10 10 530 530
%%Magnification: 1.0000
%%EndComments
270 270 translate
/Helvetica-BoldOblique 50 selectfont
350 -10 0 { gsave rotate
  0 0 moveto (PostScript) dup
  gsave 0.58 setgray show grestore
  false charpath stroke
grestore} for
showpage
```

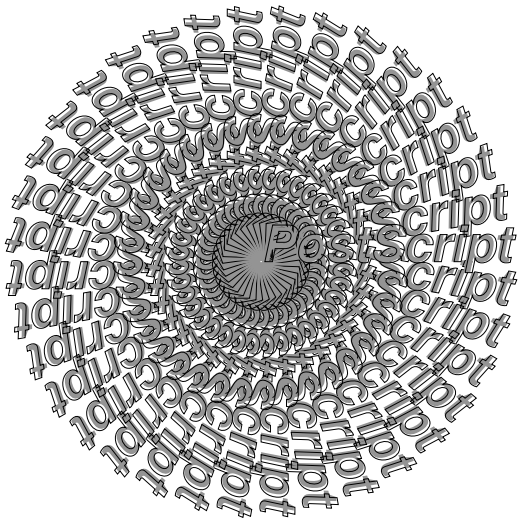

Abbildung A.2: PostScript-Beispiel

PostScript ist eine 1982 entwickelte Sprache zur Beschreibung von Druckseiten (Adobe 2000), die durch die genaue Angabe des Layouts zur Ansteuerung von Druckern gedacht war. Neben ihrer Eigenschaft als Grafiksprache kennt PostScript auch sehr viele Konstrukte höherer Programmiersprachen wie Bedingungen, Schleifen und Funktionen. Abbildung A.2 veranschaulicht dies an einem Beispiel nach Endres und Fellner (2000, S. 210)

PostScript beschreibt den Inhalt einer Seite durch geometrische Vektorfunktionen und muss daher interpretiert werden. Bei der Anzeige auf Bildschirmen übernimmt dies z. B. das Darstellungsprogramm. Viele Drucker verfügen über einen eingebauten PostScript-Interpreter und können die Dateien so direkt ausgeben. In beiden Fällen müssen aus den Vektorinformationen Rasterdaten erstellt werden, um die Geometrie der angegebenen Figuren auf dem Druck- oder Bildschirmraster darstellen zu können.

PostScript-Dateien sind textorientiert. Ihr Umfang kann teils sehr groß werden (siehe im Kolophon auf Seite 74), ist aber sehr gut komprimierbar.

Die Komplexität von PostScript war kein Hinderungsgrund, dass es sich im Grafik- und Druckbereich als Standard durchsetzte, genügte aber nicht dem Anspruch auf leichte Weitergabe von Druck- oder Layoutvorlagen.

Ebenfalls aus der Feder von Adobe stammt der PostScript-Abkömmling *Portable Document Format* (Adobe 2001). Wie schon der Name andeutet, liegt das Hauptaugenmerk auf der

[4] http://www.w3.org/StyleSheets/Core/preview/

Portabilität. PDF-Dateien sind wesentlich kleiner als PostScript-Dateien und können schon bei der Erstellung komprimiert werden.

Mit PDF hat Adobe auch eine web-fähige PostScript-Version geschaffen, die alle wesentlichen Anforderungen an eine layout-getreue Darstellung auf unterschiedlichen Plattformen garantiert. Da das Anzeigeprogramm *Acrobat Reader* kostenlos zur Verfügung steht, hat sich PDF zum Standard des elektronischen Dokumentenaustausches entwickelt und wird von Adobe gerne als *„e-Paper“* bezeichnet.

A.3 Grafikformate

Für die Grafikformate sollen hier stellvertretend zwei aufgrund ihrer Bedeutung im World Wide Web genannt werden: PNG und JPEG.

PNG (*Portable Network Graphic*) ist eines der neueren Bildformate und wurde als Ersatz für das durch Patente umstritten gewordene *Graphics Interchange Format* (GIF)[5] entwickelt und unterliegt der Obhut des W3C[6]. Es beherrscht die Speicherung von Transparenz (wie GIF) und Gammawerten, um die Farbdarstellung auf unterschiedlichen Ausgabegeräten korrigieren zu können. PNG-Bilder können auch im *Interlacing*-Modus geladen und angezeigt werden, was vor allem bei großen Bildern vorteilhaft ist: das Bild wird progressiv geladen und aus den jeweils verfügbaren Bilddaten wird eine Grobschemaanzeige erstellt, die mit den weiteren folgenden Daten immer weiter bis zur Vollständigkeit verfeinert wird. Zudem ist PNG verlustfrei komprimierbar und kann Farben mit 48 Bit pro Pixel darstellen. GIF beherrscht hier nur 8 Bits.

Das am weitesten verbreitete Format für qualitativ hochwertige Bilder ist immer noch **JPEG** (Joint Photographic Expert Group), das als ISO-Standard vorliegt (ISO 2000). In letzter Zeit leidet JPEG aber an den gleichen Problemen wie seinerzeit GIF, indem es nämlich zum Streitpunkt um Patente an enthaltenen Verfahren wird (Trinkwalder 2002).

JPEG erlaubt eine Kodierung von 24 Bit Farbdaten pro Pixel und einen variierbaren Kompressionsgrad, der allerdings verlustbehaftet ist. Eine hohe Bildqualität wird daher nur bei geringer Kompression erreicht, da ansonsten im Bild sogenannte „Artefakte“ zu sehen sind.

Gerade die Problematik von Patentrechten unterliegenden Quasi-Standards, die durch ihre weite Verbreitung nahezu unersetzlich geworden sind, zeigt wie wichtig es ist, derartige Aspekte zu beachten und zu berücksichtigen.

[5] Zur Patentproblematik sei auf König und Hüskes (1995) verwiesen.

[6] `http://www.w3.org/Graphics/PNG/`

Anhang B

Bildretrieval

B.1 Prinzip

Das *GNU Image Finding Tool* (GIFT)[1] ist eine Indizierungs- und Retrieval-Engine zur inhaltsbasierten Bildersuche (*Content Based Image Retrieval*, CBIRS), bei der versucht wird, Bilddateien anhand ihres Inhaltes und nicht durch textuelle Metadaten wie Dateinamen, Beschreibungen, Titel, etc. in einer Sammlung zu finden.

Das Prinzip der inhaltsbasierten Suche ist in den Artikeln von Müller (2001) und Ehrmann (2000) ausführlich erklärt. Die Installation und Funktionsweise von GIFT wird bei Müller (2002a) praktisch dargestellt, Müller (2002b) geht auf die Nutzung der verfügbaren Schnittstellen und die Integration in bestehende Retrieval- und weitere Systeme ein.

GIFT selbst stellt lediglich die Routinen zur Indexierung von Sammlungen und einen über MRML[2] steuerbaren Server für den Zugriff zur Verfügung. Zu dessen Nutzung ist ein entsprechender *Client* notwendig, Hinweise darauf finden sich auf den GIFT-Seiten[3].

B.2 Beispiel

Die Universität Genf, Urheberin des GIFT, stellt auf ihren Webseiten das Interface *Viper* inklusive Demo zur Verfügung.[4]

Hinter *Viper* steckt ein PHP-basierter MRML-Client, der sich um die Kommunikation mit einem GIFT-Server kümmert, aber auch jeden MRML-fähigen sonstigen Server steuern kann.

Zur Bildsuche kann man hier ein Beispielbild aus einer Zufallsmenge auswählen oder ein eigenes Bild hochladen (*„query by example“*).

Im Beispiel aus Abbildung B.1 auf der nächsten Seite wurde eine verkleinerte Fassung des Originalbildes als Beispiel hochgeladen. Das Original-Bild findet sich als erstes in der Ergebnismenge mit einer Übereinstimmung von 0,769888 zum Beispiel-Bild.

Über eine Bewertung der Ergebnisbilder (*„relevance feedback“*) lässt sich das Suchergebnis durch die Relevanz seiner Elemente weiter verfeinern.

[1] `http://www.gnu.org/software/gift/gift.html`

[2] *Multimedia Retrieval Markup Language* ⟨`http://www.mrml.net`⟩.

[3] `http://www.gnu.org/software/gift/mrml-clients.html`

[4] `http://viper.unige.ch/`

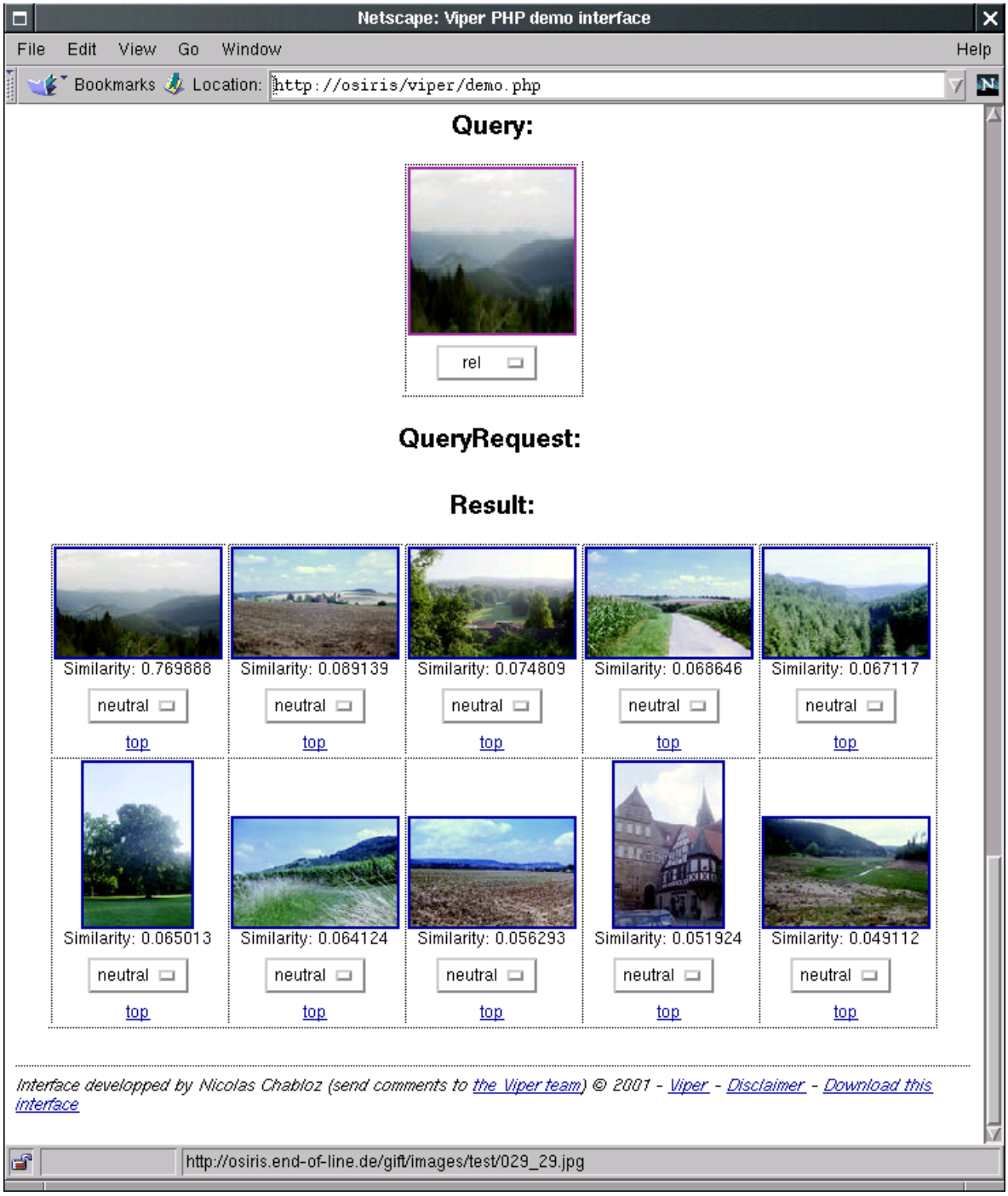

Abbildung B.1: Das Viper-Interface mit Beispielsuche

B.3 Resumée

Schon die breite Streuung der Ergebnismenge zum Beispielbild zeigt, dass es sich hier um ein aufwändiges und daher auch schwieriges Verfahren handelt. Das vorstechende Charakteristikum aller Ergebnisbilder ist zum Beispiel der helle Himmel.

Da eine Suche so nur nach sehr markanten Merkmalen erfolgen kann, werden die Grenzen auch bald deutlich. So ist es nahezu unmöglich, „alle Bilder mit Xyz darauf“ zu finden, etwa einem bestimmten Bauwerk. Dieses ist nur auffindbar, wenn die Bilder in der Sammlung möglichst viele Charakteristika mit denen des Beispiels gemeinsam haben.

Allein der große Unterschied des Originalbildes zu seiner verkleinerten Form, die als Beispiel der Abfrage diente, zeigt die Anfälligkeit dieses Verfahrens.

Anhang C

CD-ROM

Dieser Arbeit liegt eine CD-ROM im Joliet-erweiterten Format bei, welche die wichtigsten Anwendungen und Beispiele sowie die PDF- und PostScript-Dateien des Textes enthält.

`gsdl_bundle` enthält sämtliche Original-Softwarepakete und die Dokumentation der Greenstone-Software.

`gsdl_da` enthält den Installationsbaum der Greenstone-Software mit allen in dieser Arbeit beschriebenen Anpassungen und Dokumenten.

`exportet_diplom` enthält die für MicroSoft Windows exportierte Sammlung *„diplom“*.

`druck` enthält die PDF- und PostScript-Dateien dieser Arbeit.

Literaturverzeichnis

Adobe 2000 Adobe Systems Inc.: *PostScript*. 2000. – URL `http://partners.adobe.com/asn/developer/technotes/postscript.html`. – Zugriffsdatum: 2001-03-27

Adobe 2001 Adobe Systems Inc.: *Technical notes : Acrobat/PDF*. 2001. – URL `http://partners.adobe.com/asn/developer/technotes/acrobatpdf.html`. – Zugriffsdatum: 2001-03-27

ANSI 1995 American National Standards Institute: *ANSI/NISO Z39.50-1995 (version 3) Information Service Retrieval Protocol*. New York : American National Standards Institute, 1995. – 180 S. – URL `http://www.niso.org/standards/resources/Z39-50.pdf`. – Zugriffsdatum: 2002-09-07. – ISBN 1-880124-22-X

Arms 2000 Arms, William Y.: *Digital Libraries*. Cambridge, Mass. : MIT Press, 2000. – 287 S. – ISBN 0-262-01880-8

Ashley u. a. 2000 Ashley, Mike u. a.: *Das GNU-Handbuch zum Schutze der Privatsphäre*. 2000. – URL `http://www.gnupg.org/de/docs.html`. – Zugriffsdatum: 2002-09-08

Bainbridge u. a. 2001 Bainbridge, David u. a.: Greenstone: a platform for distributed digital library applications. In: *ECDL 2001*. Berlin : Springer, 2001 (Lecture notes in computer science 2163), S. 137 ff.

BDB 1994 Bundesvereinigung Deutscher Bibliotheksverbände (Hrsg.): *Bibliotheken '93 : Strukturen, Aufgaben, Positionen*. Berlin : Deutsches Bibliotheksinstitut, 1994. – 182 S. – ISBN 3-87068-445-3

Becker 2001 Becker, Peter: Wundersame Werbe-Welt : die Reklamebranche sucht neue Wege im Web. In: *c't : Magazin für Computertechnik* (2001), Nr. 19, S. 170. – ISSN 0724-8679

Berners-Lee und Fischetti 1999 Berners-Lee, Tim ; Fischetti, Mark: *Der Web-Report : der Schöpfer des World Wide Web über das Grenzenlose Potential des Internets*. München : Econ, 1999. – 332 S. – ISBN 3-430-11468-3

Berners-Lee u. a. 2001 Berners-Lee, Tim ; Handler, James ; Lassila, Ora: The Semnatic Web : a new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. In: *Scientific American* (2001), May. – URL `http://www.scientificamerican.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21&catID=2`. – Zugriffsdatum: 2002-09-08

Borenstein und Freed 1993 Borenstein, Nathaniel S. ; Freed, Ned ; Internet Engineering Task Force / Network Working Group (Hrsg.): *MIME (Multipurpose Internet Mail Extensions) Part One: Mechanisms for Specifying and Describing the Format of Internet Message Bodies*. 1993. – URL `http://www.ietf.org/rfc/rfc1341.txt`. – Zugriffsdatum: 2002-09-07

Bush 1945 Bush, Vannevar: As we may think. In: *Atlantic Monthly* 176 (1945), S. 101 – 108. – URL `http://www.theatlantic.com/unbound/flashbk/computer/bushf.htm`. – Zugriffsdatum: 2000-10-05

Callas u. a. 1998 Callas, Jon ; Donnerhacke, Lutz ; Finney, Hal ; Thayer, Rodney ; Internet Engineering Task Force / Network Working Group (Hrsg.) ; Internet Society (Hrsg.): *OpenPGP Message Format*. 1998. – URL `http://www.ietf.org/rfc/rfc2440.txt`. – Zugriffsdatum: 2002-09-08

DBI 1982 u. ö. Deutsches Bibliotheksinstitut (Hrsg.): *Regeln für den Schlagwortkatalog (RSWK)*. Berlin, 1982 u. ö.

Deutschland 1949 Deutschland, Bundesrepublik: *Grundgesetz für die Bundesrepublik Deutschland*. 1949. – Ausgabe Januar 1994

Dublin Core 1999 Dublin Core Metadate Initiative: *Dublin Core metadata element set, Version 1.1 : reference description*. 1999. – URL `http://www.dublincore.org/documents/dces/`. – Zugriffsdatum: 2001-04-10

Dublin Core 2000 Dublin Core Metadate Initiative: *Dublin Core Qualifiers*. 2000. – URL `http://www.dublincore.org/documents/dcmes-qualifiers/`. – Zugriffsdatum: 2002-09-15

Ehrmann 2000 Ehrmann, Stephan: Die Nadel im Bytehaufen : Text Retrieval, Multimediadatenbanken, Dokumentenmanagement. In: *c't : Magazin für Computertechnik* (2000), Nr. 20, S. 166 – ??. – ISSN 0724-8679

Endres und Fellner 2000 Endres, Albert ; Fellner, Dieter W.: *Digitale Bibliotheken : Informatik-Lösungen für globale Wissensmärkte*. 1. Aufl. Heidelberg : dpunkt-Verl., 2000. – 494 S. – ISBN 3-932588-77-0

Engelbart 1962 Engelbart, Douglas C.: *Augmenting Human Intellect*. 1962. – URL `http://www.histech.rwth-aachen.de/www/quellen/engelbart/ahi62index.html`. – Zugriffsdatum: 2002-08-22

Engster u. a. 2001 Engster, Florian ; Klimek, Markus ; Zimmel, Daniel: Konzeption und Aufbau der digitalen Bibliothek Information und Medien : Projektbericht des Seminars „Digitale Bibliothek“ / HdM Stuttgart, Hochschule der Medien. URL `http://diana.iuk.hdm-stuttgart.de/digbib/publ/projektbericht.pdf`. – Zugriffsdatum: 2002-09-07, September 2001. – Forschungsbericht

Ewert und Umstätter 1997 Ewert, Gisela ; Umstätter, Walter: *Lehrbuch der Bibliotheksverwaltung*. Stuttgart : Hiersemann, 1997. – 204 S. – ISBN 3-7772-9730-5

Frank u. a. 1999 Frank, Eibe u. a.: Domain-specific keyphrase extraction. In: *Proc. 16th Joint Conference on Artificial Intelligence*. San Francisco : Morgan Kaufman, 1999, S. 668 – 673

Freed und Borenstein 1996 Freed, Ned ; Borenstein, Nathaniel S. ; Internet Engineering Task Force / Network Working Group (Hrsg.): *Multipurpose Internet Mail Extensions (MIME) Part Two: Media Types*. 1996. – URL `http://www.ietf.org/rfc/rfc2046.txt`. – Zugriffsdatum: 2002-09-07

Friedling 2001 Friedling, Erik: *Struktur- und Entwicklungsplan für die Bibliothek der HdM*. September 2001. – unveröffentlichtes Positions- und Arbeitspapier

Gaus 2000 Gaus, Wilhelm: *Dokumentations- und Ordnungslehre : Theorie und Praxis des Information Retrieval*. 3., akt. Aufl. Berlin : Springer, 2000. – 452 S. – ISBN 3-540-66946-9

Geppert und Roßnagel 1998 Geppert, Martin ; Rossnagel, Alexander: *Telemediarecht : Telekommunikations- und Multimediarecht*. 1. Aufl. München : Deutscher Taschenbuch-Verl., 1998. – 320 S. – Stand: 1. Januar 1998. – ISBN 3-423-05598-7

Goodman 1987 Goodman, H. J. A.: The „World Brain/World Encyclopedia“ concept : its historical roots and the contributions of H. J. A. Goodman to the ongoing evolution and implementation of the concept. In: Chen, Ching-chih (Hrsg.): *ASIS '87 : proceedings of the 50th ASIS annual meeting* Bd. 24. Medford : Learned Information, 1987, S. 91 – 98

Gundry 2001 Gundry, John: *Knowledge Management*. 2001. – URL `http://www.knowab.co.uk/kma.html`. – Zugriffsdatum: 2002-08-08

Hacker 1992 Hacker, Rupert: *Bibliothekarisches Grundwissen*. 6., völl. neu bearb. Aufl. München : Saur, 1992. – 406 S. – ISBN 3-598-11078-2

Haun 2000 Haun, Matthias (Hrsg.): *Wissensbasierte Systeme : eine praxisorientierte Einführung*. Renningen-Malmsheim : expert-Verl., 2000. – 285 S. – ISBN 3-8169-1677-5

Henze 1999 Henze, Volker: Langzeitarchivierung elektronischer Publikationen. In: Litterski, Bärbel (Hrsg.) ; Harder, Uwe (Hrsg.): *Bücher, Bytes und Bibliotheken : Integrierte Information im Internet. 4. InetBib-Tagung vom 3. – 6. Märt 1999 in Oldenburg*. Dortmund : Universitätsbibliothek Dortmund, 1999, S. 86 – 87. – ISBN 3-921823-25-X

ISO 1986 International Organisation for Standardization: *Information processing - Text and office systems - Standard Generalized Markup Language (SGML)*. 1. ed. Geneva : ISO, 1986. – XI, 155 S. S. – Norm-Nr.: ISO 8879-1986

ISO 2000 International Organisation for Standardization: *Information technology – JPEG 2000 image coding system*. Geneva : ISO, 2000. – Norm-Nr.: ISO/IEC 15444:2000

Jochum 1993 Jochum, Uwe: *Kleine Bibliotheksgeschichte*. Stuttgart : Reclam, 1993. – 232 S. – ISBN 3-15-008915-8

Jones 1998 Jones, Steve: *Link as you type : using key phrases for automated dynamic link generation*. 1998. – URL `http://www.cs.waikato.ac.nz/~nzdl/publications/1998/Jones-LinkAsYouType.pdf`. – Zugriffsdatum: 2002-09-07

Jones u. a. 1999 Jones, Steve ; McInnes, S. ; Staveley, Mark S.: A graphical user interface for Boolean query specification. In: *International Journal on Digital Libraries* (1999), Nr. 2, S. 207 – 223

Jones und Staveley 1999 Jones, Steve ; Staveley, Mark S.: Phrasier : a system for interactive document retrieval using keyphrases. In: Hearst, M. (Hrsg.) ; Gey, F. (Hrsg.) ; Tong, R. (Hrsg.): *Proceedings of the 22. International Conference on Research and Development in Information Retrieval (SIGIR '99)*. Berkeley, August 1999, S. 160 – 167

Klatt u. a. 2001 Klatt, Rüdiger u. a.: *Nutzung elektronischer wissenschaftlicher Information in der Hochschulausbildung*. URL `http://www.stefi.de/`. – Zugriffsdatum: 2001-07-15, 2001

König und Hüskes 1995 König, Volker ; Hüskes, Ralf: Lizenzquerelen : Lizenzgebühren für GIF erhoben. In: *c't : Magazin für Computertechnik* (1995), Nr. 3, S. 29. – ISSN 0724-8679

Lem 1997 Lem, Stanislaw: *Exformation : die explosive Information*. 1997. – URL `http://www.telepolis.de/deutsch/kolumnen/lem/2108/1.html`. – Zugriffsdatum: 2002-03-17

Lobin 2000 Lobin, Henning: *Informationsmodellierung in XML und SGML*. Berlin : Springer, 2000. – 234 S. – ISBN 3-540-65356-2

Maile und Scholze 1997 Maile, Annette ; Scholze, Frank: *Online Publikationsverbund der Universität Stuttgart (OPUS)*. 1997. – URL `http://elib.uni-stuttgart.de/opus/volltexte/1999/226/`. – Zugriffsdatum: 2002-09-15

McKnight 1997 McKnight, Cliff: *Electronic Library*. S. 130 – 132. In: Feather, John (Hrsg.) ; Sturges, Paul (Hrsg.): *International Encyclopedia of Information and Library Science*. London : Routledge, 1997

McNab u. a. 1996 McNab, Roger J. u. a.: Towards the digital music library : tune retrieval from acoustic input. In: Fox, E. A. (Hrsg.) ; Marchionini, G. (Hrsg.): *Proc. Digital Libraries '96*. New York : ACM Press, 1996, S. 11 – 18

Meyer 2002 Meyer, Angela: Linkfäule : Wissenschaftler untersuchen die Lebensdauer von Hyperlinks. In: *c't : Magazin für Computertechnik* (2002), Nr. 9, S. 54. – ISSN 0724-8679

Müller 2001 Müller, Henning: Suchen ohne Worte : wie inhaltsbasierte Suche funktioniert. In: *c't : Magazin für Computertechnik* (2001), Nr. 15, S. 162 ff.. – ISSN 0724-8679

Müller 2002a Müller, Henning: Jäger des verlorenen Fotos : das GNU Image Finding Tool für Linux in der Praxis. In: *c't : Magazin für Computertechnik* (2002), Nr. 6, S. 252 – 257. – ISSN 0724-8679

Müller 2002b Müller, Wolfgang: Bildersuchbaukasten : das GNU Image Finding Tool via Plug-ins erweitern. In: *c't : Magazin für Computertechnik* (2002), Nr. 17, S. 190 – 195. – ISSN 0724-8679

Nelson 1992 Nelson, Theodor H.: *Literary machines : the report on, and of, project Xanadu concerning word processing, electronic publishing, hypertext, thinkertoys, tomorrow's intellectual revolution, and certain other topics including knowledge, education and freedom.* Ed. 93.1. Sausalito : Mindfull Press, 1992

Paynter und Witten 2001 Paynter, Gordon W. ; Witten, Ian H.: *A combined phrase an thesaurus browser for large document collections.* 2001. – URL `http://www.scils.rutgers.edu/~nina/phrasebrowsing/workshop062801/WittenPaper.pdf`. – Zugriffsdatum: 2001-10-05

Pitschmann 2001 Pitschmann, Louis A.: *Building sustainable collections of free third-party web resources.* Washington, D.C. : Digital Library Federation, 2001. – VI, 44 S. – URL `http://www.clir.org/pubs/reports/pub98/pub98.pdf`. – Zugriffsdatum: 2002-05-25. – ISBN 1-887334-83-1

Rehm 1991 Rehm, Margarete: *Lexikon Buch, Bibliothek, Neue Medien.* München : Saur, 1991. – 294 S. – ISBN 3-598-10851-6

Rivest und RSA Data Security 1992 Rivest, Ronald L. ; RSA Data Security ; Internet Engineering Task Force / Network Working Group (Hrsg.): *The MD5 Message-Digest Algorithm.* 1992. – URL `http://www.ietf.org/rfc/rfc1321.txt`. – Zugriffsdatum: 2002-09-07

Sparck-Jones 1997 Sparck-Jones, Karen (Hrsg.): *Readings in information retrieval.* San Francisco : Morgan Kaufmann, 1997. – XV, 589 S. – ISBN 1-55860-454-5

Trinkwalder 2002 Trinkwalder, Andrea: Lizenzpoker : Texanisches Unternehmen will für JPEG kassieren. In: *c't : Magazin für Computertechnik* (2002), Nr. 17, S. 24. – ISSN 0724-8679

UNESCO 2000 UNESCO: Digital Libraries. In: *UNESCO Sources* June 2000 (2000), Nr. 124, S. 12. – ISSN 1014-6989

Weeks 2001 Weeks, Linton: Pat Schroeder's new chapter : the former congresswoman is battling for america's publishers. In: *The Washington Post* (2001), February 7, S. C01. – URL `http://www.washingtonpost.com/wp-dyn/articles/A36584-2001Feb7.html`. – Zugriffsdatum: 2002-09-10

Wells 1938 Wells, Herbert G.: *World Brain.* London : Methuen, 1938. – 130 S

Wissenschaftsrat 2001 Wissenschaftsrat: *Empfehlung zur digitalen Informationsversorgung durch Hochschulbibliotheken.* 2001. – URL `http://www.wissenschaftsrat.de/texte/4935-01.pdf`. – Zugriffsdatum: 2001-07-15

Witten u. a. 1999a Witten, Ian H. u. a.: KEA: Practical automatic keyphrase extraction. In: *Proc. DL '99.* New York : ACM Press, 1999, S. 254 – 256

Witten und Boddie 2002a Witten, Ian H. ; Boddie, Stefan: *Greenstone Digital Library Developer's Guide.* 2002. – URL `http://www.greenstone.org/englisch/docs.html`. – Zugriffsdatum: 2002-09-08

Witten und Boddie 2002b WITTEN, Ian H. ; BODDIE, Stefan: *Greenstone Digital Library Installer's Guide*. 2002. – URL `http://www.greenstone.org/englisch/docs.html`. – Zugriffsdatum: 2002-09-08

Witten und Boddie 2002c WITTEN, Ian H. ; BODDIE, Stefan: *Greenstone Digital Library User's Guide*. 2002. – URL `http://www.greenstone.org/englisch/docs.html`. – Zugriffsdatum: 2002-09-08

Witten u. a. 1999b WITTEN, Ian H. ; MOFFAT, Alistair ; BELL, James: *Managing gigabytes : compressing and indexing documents and images*. 2nd ed. San Francisco : Morgan Kaufmann, 1999. – XXXI, 519 S. – ISBN 1-55860-570-3

Wolf 1995 WOLF, Gary: The Curse of Xanadu. In: *Wired* (1995), Nr. 6. – URL `http://www.wired.com/wired/archive/3.06/xanadu.html`. – Zugriffsdatum: 2002-09-13

Yeates 1999 YEATES, Stuart A.: *Novel indexes and metadata sources in digital libraries*, University of Waikato, Dissertation, August 1999. – 35 S

Zivadinovic 2001 ŽIVADINOVIĆ, Dušan: Warten auf schnelle Analogmodems. In: *c't : Magazin für Computertechnik* (2001), Nr. 3, S. 31. – ISSN 0724-8679

Kolophon

Dieser Text wurde gesetzt mit LaTeX 2_ε aus der 11 Punkt Computer-Modern-Schrift. Für den Schriftsatz wurden die ec-Schriften von Jörg Knappen verwendet. Die PostScript- und PDF-Dateien dieses Dokumentes verwenden die Type-1-Varianten `cm-super` von Vladimir Volovich ⟨vvv@vsu.ru⟩.

Das Literaturverzeichnis wurde nach DIN 1505 mit BibTeX in der Version 0.99c und dem `dinat`-Paket in der Version 2.5 von Helge Baumann ⟨helge.baumann@gmx.de⟩ erstellt.

Zur vollständigen Auflösung aller Referenzen benötigten LaTeX und seine Hilfsprogramme insgesamt 5 Läufe.

Der Satz erfolgte auf einer unter GNU/Linux betriebenen Pentium-Maschine mit 233 Megahertz Taktung unter Verwendung der teTeX-Distribution in der Version 1.0.7. Diese baut auf Web2C, Version 7.3.7, kpathsea, Version 3.3.7 und TeX/LaTeX, Version 3.14159 auf.

Alle Eingabe- und Steuerdateien brachten eine Summe von 2,1 Megabytes auf. Die Größe der Ausgabedateien betrug im DVI-Format 228 Kilobytes (ohne Grafiken), im PS-Format 30 Megabytes und im (aus Qualitätsgründen unkomprimierten) PDF-Format 15 Megabytes.

Erklärung

Hiermit erkläre ich, dass ich die vorliegende Diplomarbeit selbstständig angefertigt habe. Es wurden nur die in der Arbeit ausdrücklich genannten Quellen und Hilfsmittel benutzt. Wörtlich oder sinngemäß übernommenes Gedankengut habe ich als solches kenntlich gemacht.

Ort und Datum Unterschrift